nature

www.nature.com/nature

Vol 462 | Issue no. 7275 | 17 December 2009

# Credit where credit is due

A proposed author ID system is gaining widespread support, and could help lay the foundation for an academic-reward system less heavily tied to publications and citations.

In his classic book *Management Teams*, UK psychologist Meredith Belbin used extensive empirical evidence to argue that effective teams require members who can cover nine key roles. These roles range from the creative 'plants' who generate novel ideas, to the disciplined 'implementers' who turn plans into action and the big-picture 'coordinators' who keep everyone working together.

Much the same range of roles is critical for science. Unfortunately, the academic system tends to reward only some of those activities — notably those that have easily measured outcomes, such as the publication and citation numbers so heavily weighted by promotion and tenure committees.

On page 843, *Nature* profiles a research group composed largely of what Belbin would call 'completer finishers' — perfectionists who are driven to fix all the flaws, fill all the gaps and get the job finished correctly. This particular group is trying to complete the reference human genome sequence, which is still full of errors nearly a decade after the first draft was announced in 2000. It is essential work: modern sequencing techniques still use the reference to anchor new data even as they grind out genomes at a fraction of the cost of the original. But it is also work that offers few academic rewards beyond the satisfaction of a job well done — it is unlikely to result in a high-profile publication.

Such unsung contributions to science may soon be easier to evaluate and quantify through an author ID system proposed earlier this month and backed by 23 organizations, including Thomson Reuters, Nature Publishing Group, Elsevier, ProQuest, Springer, CrossRef, the British Library and the Wellcome Trust. The Open Researcher and Contributor ID (ORCID) would be an alphanumeric string that uniquely identifies an individual scientist in much the same way that a Digital Object Identifier uniquely identifies a paper, book or other scholarly publication (more details and the complete list of participants will soon be available at www.orcid.org). The system would distinguish between the world's multitudinous Dr Smiths and Professor Wangs, but would not be affected by name changes, cultural differences in name order, inconsistent first-name abbreviations or the use of different alphabets. It would be attached to researchers' journal publications, and could also be assigned to data sets they helped to generate, comments on their colleagues' blog posts or unpublished draft papers, edits of Wikipedia entries and much else besides.

This kind of 'microattribution' could ultimately make it possible for each researcher to have a constantly updated 'digital curriculum vitae' providing a picture of his or her contributions to science going far beyond the simple publication list.

ORCID is hardly the first proposal for an author ID system. A number of publishers have been exploring the idea, and the International Organization for Standardization in Geneva is developing an international standard name identifier to track contributors to media content such as books, television programmes and newspaper articles. But most of those developers have already joined or are working closely with the ORCID group. Moreover, the intention is to make ORCID freely available for anyone to use, and interoperable with existing ID systems.

**"The ID system could make it possible for each researcher to have a constantly updated 'digital curriculum vitae'."**

The next step is for the ORCID group to turn the concept into a working system. That is scheduled to happen over the next six months, with the software being based on Thomson Reuters' existing ResearcherID system. In parallel, the group will be setting up an independent organization to run the system and assign ORCIDs to individual researchers.

There will be many challenges along the way — not least of which is establishing rigorous protocols for validating and authenticating ORCID assignments. No one wants to see the system abused by individuals seeking to pad their academic credentials.

But perhaps the largest challenge will be cultural. Whether ORCID or some other author ID system becomes the accepted standard, the new metrics made possible will need to be taken seriously by everyone involved in the academic-reward system — funding agencies, university administrations, and promotion and tenure committees. Every role in science should be recognized and rewarded, not just those that produce high-profile publications. ■

# Mind the gap

It will take time to assess the value of fresh approaches to science and technology studies.

The relationship between the social sciences and the natural sciences has historically been fraught. 'Hard' scientists have often treated the social sciences with disdain. For example, some of them fought, successfully at first, to exclude the social sciences from the remit of the US National Science Foundation. And those social scientists who studied science itself, under the remit of science and technology studies, often returned the favour, seeming on occasion to be devoting themselves myopically to demonstrating that the scientific emperor had few, if any, clothes.

There remains something of a dialogue of the deaf between these two wings of the academy, separated as they are by language, custom and methodology. But barriers are coming down. Senior scientists and administrators, especially those in socially contentious areas such as climate change and reproductive technologies, realize that

they need to collaborate with scholars of society-at-large. Sociologists and philosophers of science, in turn, are acquiring a more intimate understanding of the scientists that they study.

These promising developments are being driven by a wider political context. In the United States, the events of 11 September 2001 and the subsequent wars in Afghanistan and Iraq led to a reassessment of the role of social-sciences research, particularly in regard to its relevance to national security. This led to firmer, bipartisan support for the social, behavioural and economic sciences directorate at the National Science Foundation. In Europe, meanwhile, strong public suspicion of new technologies — which has had particularly devastating consequences for the deployment of genetically modified crops in Europe and beyond — has encouraged governments to set aside more resources for the early involvement of social scientists in technology development.

In this issue, we report on the strengths and weaknesses of a UK initiative to bring the social sciences to bear more effectively on genomics, and on the life sciences more generally (see page 840). This experiment suggests that coherent, multidisciplinary centres can help social scientists to get a firmer grip on the complex science, cultures and behaviours underlying new technologies. But it also highlights the need for funding agencies, such as the UK Economic and Social Research Council, to retain a close interest in the strategic direction of such centres, and to ensure that their successes and failures are noted and built upon, even after their direct funding has expired.

The increased involvement of social scientists in science and technology issues has been especially pronounced of late in nascent fields such as nanotechnology and synthetic biology, where funding agencies feel that they have to tread carefully lest their work unleashes a backlash from the public.

In these areas, it is too early to assess the value — and beneficiaries — of the social scientists' contribution. The idea of embedding sociology, law and philosophy firmly in the development of a scientific discipline from the outset is only now being tested. There is optimism among many of the engineers and natural and social scientists involved.

However, all the signs are that the various parties are approaching these collaborations very much on their own terms. Natural scientists are under pressure to deliver new insight and applications. As far as they are concerned, if social scientists wish to observe them, that's probably tolerable.

Social scientists, in turn, wish to be respected for their insight into how scientists and their ideas function both within their communities and, above all, in relation to societal ambitions and values. These researchers do not, by and large, see their purpose as being to pre-empt societal reactions or public engagement, or to help natural scientists communicate. Moreover, for them, the tension between collaboration and detachment in these projects is real.

**"Coherent, multidisciplinary centres can help social scientists to get a firmer grip on the complex science, cultures and behaviours underlying new technologies."**

There is a possibility, therefore, that these parties will end up walking and talking past each other. What is more, the management of rigorous programmes involving both groups is hampered by the difficulty that the social scientists (and those who support them) have in reaching agreement on what constitutes outstanding analysis of human practices in these contexts: just how that can best be achieved, and to what extent, should be at the service of government policy goals.

None of this should encourage a dismissive attitude among sceptics. The applications of genetics, nanotechnology, synthetic biology and other technologies are giving rise to substantial new challenges in professional practice and communication, in ethics, in intellectual property and in many other dimensions beyond the science itself. Objective insights into these dimensions have their own value, and the new collaborations should help. The challenge remains to identify how that value can best be fulfilled. ■

# A class of their own

## The Japanese winners of *Nature*'s mentoring awards have the universal qualities of outstanding advisers.

"Dr Kitano is always ready to invest in apparently absurd ideas. … He actively seeks to gain international exposure for his young researchers by making them corresponding authors on his papers. … His distinctive mentoring style — decisiveness and respect for the individual — comes from the fact that he is not a pure product of the Japanese system."

"The thing that most surprised me when I joined the Oosawa group was that everybody called him 'Oosawa-*san*' [rather than the much more formal 'Oosawa-*sensei*']. Also, rather than sitting in an office … he walked around the lab collaring people and talking with them."

These two extracts are drawn from the enthusiastic nominations of Hiroaki Kitano, head of Sony Computer Science Laboratories in Tokyo, and Fumio Oosawa, a biophysicist at Aichi Institute of Technology in Toyota — the respective winners of the 2009 'mid career' and 'lifetime achievement' awards given by *Nature* for scientific mentoring. Since the awards' inception in 2005, they have been held in a different country every year, and they have been judged each time by a multidisciplinary panel of leading scientists from that country (see go.nature.com/Rccbo4).

An account of this year's awards, which took place in Japan, can be found on page 948. As in previous years, the two winners display accessibility, a broad and insightful overview, and an ability to engage with young researchers on the latter's own terms — qualities that seem to be common to outstanding mentors everywhere. There is no doubt that the Japanese system tends to be strongly hierarchical, but it was clear to the judges that, as in all other countries in which the competition has been held, qualities that buck such hierarchies lead to outstanding new generations. Congratulations to Oosawa and Kitano. ■

# RESEARCH HIGHLIGHTS

## Monkey talk

*Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0908118106 (2009)

The vocal communication system used by Campbell's monkeys may represent the most complex syntax-like structure yet found among animals.

Karim Ouattara and Alban Lemasson of the University of Rennes in France and Klaus Zuberbühler of the University of St Andrews, UK, recorded and analysed the calls of males in six groups of free-ranging Campbell's monkeys in the rainforest of Ivory Coast.

The males have just six basic types of call, but combine these in context-specific sequences to convey different information. Crowned eagles, for example, elicited four different sequences, and leopards three, according to how the male learnt about their presence — by seeing them, hearing them, or learning about them through the hearsay of other monkey species.

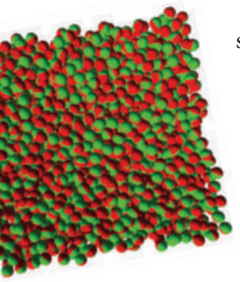
F. MÖLLERS/TAI MONKEY PROJECT

PHYSICAL CHEMISTRY

## Dual-aspect particles

*Phys. Rev. Lett.* **103,** 237801 (2009)

Like their mythological namesake, Janus particles have two faces: one that attracts and one that repels a liquid. Scientists are interested in the nanoparticles' behaviour because they mimic that of many biological and chemical molecules.



AM. PHYS. SOC.

When suspended in solution, the repellent faces cluster together, causing the particles to clump. Francesco Sciortino of the University of Rome La Sapienza and his colleagues have now found that this clumping affects gas-to-liquid phase transitions of the particles. Simulations of up to 5,000 particles in solution show that the clumping creates unusual behaviour: contrary to expectation, the gas phase (pictured above) is more ordered than the liquid phase (pictured right) and the material expands as it cools.

The researchers believe that their simulations could prompt new experimental work with Janus particles.

ECOLOGY

## Reef regulation

*Proc. R. Soc. B* doi:10.1098/rspb.2009.2062 (2009)

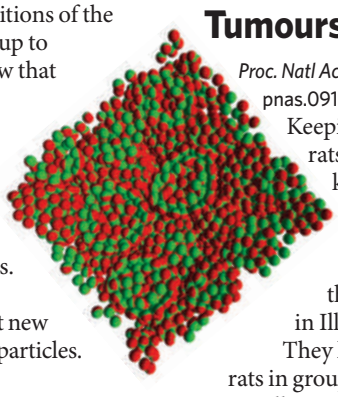
An ecosystem's stability is postulated to increase as its number of species goes up, owing to the increased number of interactions between those species and the subsequent robustness of the ecological network.

Carl Simpson and Wolfgang Kiessling of the Berlin Museum of Natural History propose an explanation for this relationship on evolutionary timescales. They say the 'diversity–stability' relationship can be explained solely by the extinction of species: high species turnover needs to be buffered by higher species numbers.

If this is true, then the diversity–stability relationship should be strongest when the extinction rate is high. Looking at historical coral reef data, they found that the relationship was historically strong during periods of high extinction, and weak during low-extinction periods.

CANCER BIOLOGY

## Tumours hate company



*Proc. Natl Acad. Sci. USA* doi:10.1073/pnas.0910753106 (2009)

Keeping normally gregarious rats isolated from their own kind boosts their cancer risk, according to Martha McClintock and her colleagues at the University of Chicago in Illinois.

They kept 20 rats alone and 20 rats in groups of five. All 40 were genetically prone to mammary cancer. The lone rats exhibited a 135% increase in the number of tumours, an 8,391% increase in the size of tumours and a 3.3-fold increase in the relative risk of malignancy compared with those kept in groups.

Isolated rats were also more stressed, anxious, fearful and vigilant. The authors suggest that prolonged exposure to large pulses of the stress-related hormone corticosterone may have contributed to tumour origin and growth.

GEOLOGY

## Bubble batholiths

*Lithosphere* **1,** 323–327 (2009)

In some mountains and plateaux, geologists find granite rocks that formed from magmas that had risen up through 'floating' continental tectonic plates. The rocks' origin has often been attributed to melting underneath the continental plates, caused by convection in the hot mantle.

However, Donna Whitney at the University of Minnesota in Minneapolis and her colleagues suggest that subduction — the downward thrusting of one plate under another during tectonic collisions — might be responsible. Using a numerical model, the researchers find that continental subduction can lead to melting of crustal slabs and percolating granitic magma.

POPULATION GENETICS

## Asia's common origin

*Science* **326,** 1541–1545 (2009)

Humans migrated from Africa into Asia, along its southern coast and then down into Indonesia. But whether this wave also accounted for east Asian populations or was supplemented by one or more later migratory waves along a northern route has been the subject of debate.

The HUGO Pan-Asian SNP Consortium reports an analysis of nearly 55,000 variations in genes from nearly 2,000 people that supports the single-wave theory. The analysis

finds a high degree of overlap between the genomes of all southeast Asians and east Asians, lesser genetic similarity with caucasian populations, and a decreasing genetic diversity from southern to northern China, suggesting that humans entered Asia in a single primary migratory wave.

CHEMISTRY

## One-hit wonder

*Nature Chem.* doi:10.1038/nchem.477 (2009)

Ammonia provides the nitrogen for most synthetic chemicals. But its industrial synthesis from nitrogen gas relies on high temperatures and pressures, and gobbles up fossil fuel.

Building on previous work on splitting the strong nitrogen–nitrogen triple bond, Paul Chirik and his colleagues at Cornell University in Ithaca, New York, have now coaxed nitrogen to react with another abundant gas, carbon monoxide, in one room-temperature step. The reaction forms carbon–carbon and carbon–nitrogen bonds, the backbones of many useful chemicals, and is orchestrated by a compound containing the rare metal hafnium. This compound is not catalytic, so is unlikely to find widespread use. But the reaction's nitrogen-weakening mechanism may inform new ways to assemble complex molecules from simple gases.

PALAEONTOLOGY

## Dawn of the anomodonts

*Proc. R. Soc. B* doi:10.1098/rspb.2009.0883 (2009)

The anomodonts were mammal-like reptiles that were widespread from 270 million years ago until at least 200 million years ago. A new specimen of an animal called *Biseridens qilianicus* has recently been unearthed in Gansu, China. The specimen is in such good shape (pictured, below) that Jun Liu of the Chinese Academy of Sciences in Beijing and his colleagues were able to confirm an earlier hunch that this animal is a very early anomodont. In fact, it is the most basal anomodont yet found, meaning that it is a member of the oldest branch on the anomodont family tree.

This analysis supports the idea that anomodonts originated on the old northern continent of Laurasia rather than on its southern counterpart, Gondwana, as previously thought.

PSYCHOLOGY

## Personality versus mood

*Arch. Gen. Psychiatry* **66,** 1322–1330 (2009)

The antidepressant paroxetine doesn't just make people happier, it alters their personality as well.

Tony Tang at Northwestern University in Evanston, Illinois, and his colleagues studied changes in neuroticism and extraversion — two personality traits linked to depression and the neurotransmitter serotonin — in 240 patients in a 16-week trial with a one-year follow-up. Half of the patients received paroxetine, one quarter a placebo and one quarter cognitive therapy.

Placebo patients improved their depression scores but reported little change in personality. By contrast, patients on paroxetine reported a decrease in neuroticism and an increase in extraversion, even after the results were normalized for differences in depression improvement. Those with the greatest declines in neuroticism also showed lower relapse rates.

Rather than being a mere by-product of improved mood, these personality changes may help explain why drugs such as paroxetine work against depression in the first place.

EPIDEMIOLOGY

## Malaria's mark

*Science* **326,** 1546–1549 (2009)

The deadliest of the four human malaria parasites, *Plasmodium falciparum*, has left its imprint on the human genome in the form of malaria-protective mutations, including those that cause sickle-cell anaemia. Now, Lluis Quintana-Murci and Anavaj Sakuntabhai at the Pasteur Institute in Paris and their colleagues show that — in a similar trade-off — pressure from a neglected strain, *P. vivax*, may maintain a common enzyme deficiency in southeast Asia that can cause jaundice and anaemia.

The team found that the local gene variant associated with the enzyme deficiency was also associated with a 30–60% reduction in parasite density of *P. vivax* but not *P. falciparum*. People with two copies of the gene had the lowest parasite densities. The results suggest that *P. vivax* has had a larger effect on the human genome than previously thought.

**Correction**

The Research Highlight 'Rude awakening' (*Nature* **462,** 547; 2009) incorrectly described the green parts of the image. The figure shows an expression pattern of green fluorescent protein (GFP) in fruitfly brains, which overlaps with expression of dopamine receptors.

## JOURNAL CLUB

**Reuben Shaw**
**The Salk Institute for Biological Studies, La Jolla, California**

**A cancer researcher ponders a fundamental connection between nutrients and gene expression.**

Nutrient availability to single-celled organisms varies according to their environment, and proteins in the cell that sense nutrient levels alter gene expression to increase uptake and use of specific metabolites to fuel cellular processes. Conversely, most cells in multicellular organisms are exposed to constant nutrient levels by the bloodstream, and so far there are few examples of metabolism being directly coupled to the control of gene expression.

A recent paper by Craig Thompson and his colleagues at the University of Pennsylvania in Philadelphia uncovers a direct connection between a well-known metabolic enzyme — ATP citrate lyase (ACL) — and changes in gene expression (K. E. Wellen *et al. Science* **324,** 1076–1080; 2009). Through a chain of reactions, ACL influences the functioning of the histones, proteins that package lengths of DNA — and unpackage them for 'reading'. This means that there is a basic — and surprising — relationship between cell glucose levels and gene expression.

We don't yet know how metabolic challenges — for example, fasting — in whole organisms affect ACL levels or activity. But we do know that some of the same proteins that increase tumour growth also modify ACL by attaching phosphorus.

It is likely that we are just at the tip of the iceberg in terms of our understanding of the molecular basis of how metabolic inputs dictate gene-expression changes in mammalian cells. Future studies using genetic models of ACL loss in distinct mouse tissues, as well as chemical inhibitors of the enzyme, will help to elucidate in which contexts it is critical for gene-expression changes in the whole organism. Moreover, our knowledge of this metabolic linchpin may provide a therapeutic window for the treatment of certain forms of cancer, almost all of which undergo metabolic adaptation.

Discuss this paper at **http://blogs.nature.com/nature/journalclub**

# NEWS BRIEFING

## POLICY

**UK space agency:** The British government has announced plans for a national space agency. The agency will replace the British National Space Centre, which was set up by the government in 1985 to coordinate space research. The new agency will consolidate the efforts of six government departments, three science and technology funding bodies and the Met Office. But it is not yet clear whether it will have independent control of funds for space activities.

**Agriculture reforms:** The Consultative Group on International Agricultural Research voted for wide-reaching organizational changes on 8 December. The public–private partnership, which oversees 15 research centres and supports some 8,000 scientists and staff across the world, has established thematic science programmes and set up a trust fund to manage US$500 million in annual donations.


B. STRONG/REUTERS

## CRUNCH TIME AT COPENHAGEN

After being derailed by protests from developing countries, the United Nations climate talks in Copenhagen are back to something resembling order. As *Nature* went to press, government ministers led by conference president Connie Hedegaard and Yvo de Boer, executive secretary of the United Nations Framework Convention on Climate Change (both pictured), were preparing a negotiating text for some 130 heads of state. But the architecture of any possible agreement was still unclear, and a deep divide remained between developed and developing nations. More than 45,000 delegates, journalists and lobbyists had registered to attend — three times the capacity of the Bella Center, where the talks are being held. For daily updates, see nature.com/roadtocopenhagen.

**Xenotransplant trials:** Australia has lifted a 5-year ban on clinical trials in which animal cells, tissues or organs are transplanted into humans. The National Health and Medical Research Council said on 10 December that it was satisfied the risks of transmitting animal viruses in this way were low, and that trials could proceed once regulatory and monitoring frameworks have been established. The United States, Russia, European Union and China are among many places that already allow human xenotransplantation trials, the council said.

**Countering bioweapons:** The United States announced on 9 December that it would try to revitalize the Biological Weapons Convention (BWC) that bans the development, production and stockpiling of such weapons. Ellen Tauscher, undersecretary of state for arms control, said that the country would urge all nations to sign the BWC, but rejected efforts to develop a system to verify that the convention was being obeyed. The treaty has foundered since the administration of former president George W. Bush broke with attempts to set up a compliance scheme in 2001 (see *Nature* **414,** 675; 2001).

**Budget threat:** British scientists were rattled by the UK government's 2010 pre-budget report on 9 December, which called for a £600-million (US$975-million) reduction in spending on higher education, science and research for the period 2011–13. Details of the cutbacks were not specified — and the present Labour government may not win next year's general election — but their mention reminded researchers that painful spending cuts almost certainly lie in wait.

## NUMBER CRUNCH

**–19%**

**The average investment return in the 2009 financial year from more than 500 US college and university endowments.**

*Source: NACUBO-Commonfund Study of Endowments*

**Drug-safety rap:** Five years after the blockbuster painkiller Vioxx was withdrawn after being linked to heart attack and stroke, the US Food and Drug Administration (FDA) is still not taking adequate steps to ensure the safety of marketed drugs, says a report by the Government Accountability Office (GAO). The report, made public on 9 December, checked the agency's progress in enacting recommendations from a 2006 GAO report. It found that the FDA had failed to transfer sufficient authority for post-market drug safety to the Office of Surveillance and Epidemiology and away from the office that approves new drugs.

## BUSINESS

**Geothermal project buried:** An innovative geothermal project that would drill deep into the Earth to extract energy from hot rocks has been shut down. On

10 December, public authorities in Basel, Switzerland, closed a private geothermal drilling project led by Geopower Basel, after a government study found that the small earthquakes it generated would cause too much economic damage. The project had been on hold since 2006. See page 848 for more.

**Solar-power boost:** The World Bank announced on 9 December that it planned US$5.6 billion of financing to speed up the deployment of concentrated solar power in North Africa and the Middle East. Its Clean Technology Fund would spend $750 million on 11 projects in Algeria, Egypt, Jordan, Morocco and Tunisia, and $4.85 billion would be mobilized from other sources, including donors and commercial debt. The investment will create around a gigawatt of solar capacity over the next five years, the World Bank says — a sizeable addition to the 6–7 gigawatts of global projects in the pipeline by the end of 2008.

## RESEARCH

**Harvard stops building:** With a hobbled endowment, Harvard University in Cambridge, Massachusetts, announced on 10 December that it would halt construction of the US$1-billion science complex on its Allston campus early next year once the foundations are complete. This brings to a standstill the university's ambitious expansion plans (see *Nature* **454**, 686–689; 2008). Instead, the university said it would shift focus towards improving and leasing vacant Allston properties. It is not clear when building work might begin again; the university is reviewing financing options.

**Protein structures removed:** The University of Alabama at Birmingham has asked for twelve entries to be removed from the Protein Data Bank. This follows the recommendation of an expert committee that has been looking into allegations of suspect data. The protein structures are cited in hundreds of papers. Last week, *The Journal of Biological Chemistry* retracted a paper containing one of the structures.

**Marine viewing:** A sea-floor observatory in the northeast Pacific Ocean officially went live on 8 December, promising free access via the Internet to its live camera feeds and to data on marine life off British Columbia. The Can$100-million (US$95-million) NEPTUNE Canada project (http://neptunecanada.ca) will study biological, physical and geological processes up to 300 kilometres offshore, using instruments and sensors connected by a cable that runs electricity and fibre-optics in an 800-kilometre undersea loop.

**Malaria optimism:** A "tremendous increase in funding" for malaria has shown positive results, the World Health Organization (WHO) said on 15 December, as it published its *World Malaria Report 2009*. US$1.7 billion was committed in 2009, compared with $730 million in 2006, allowing greater distribution of insecticide-treated bed nets and artemisinin-based therapies. From 2000 to 2008, malaria cases have been cut by at least 50% in more than one-third of malarious countries. But the WHO says that $5 billion is still required annually, and resistance to artemisinin remains a threat.

## NEWS MAKER


NASA/MOORE, VAFB

WISE

**The Wide-Field Infrared Survey Explorer, a space telescope, was launched by NASA on 14 December.**

## THE WEEK AHEAD

**17–18 DECEMBER**

**The governing council of CERN meets. Its agenda includes possible expansion of the organization's international membership — Israel and Turkey are among those who have applied to join.**

go.nature.com/jk7A9T

**18 DECEMBER**

**The United Nations Climate Change Conference in Copenhagen comes to an end.**

http://en.cop15.dk

**19 DECEMBER**

**South Korea's first icebreaker research ship, *ARAON*, is scheduled to set sail for the Antarctic on a three-month expedition.**

www.kopri.re.kr/index_eng.jsp

**Synchrotron stand-off continues:** Scientists at the Australian Synchrotron in Melbourne voted last week to again work only between 9 a.m. and 5 p.m. in renewed protest at the facility's governing board. The synchrotron is undergoing maintenance over the Christmas break, but if the limited-hours schedule continues in 2010, it will cripple research projects in a facility usually booked around-the-clock. Five members of the synchrotron's scientific advisory committee also resigned last week (see *Nature* **462,** 706–707; 2009).

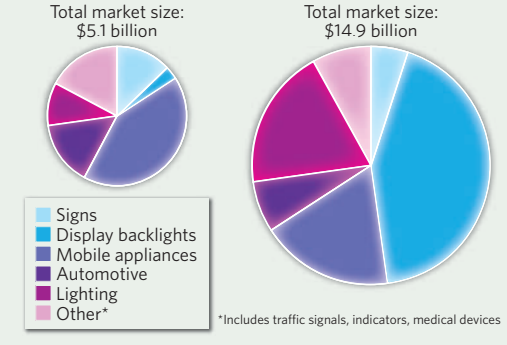
## BUSINESS WATCH

Makers of light-emitting diodes (LEDs) are preparing to increase capacity in response to growing demand. LED suppliers were operating at 30% of capacity earlier this year, but are now producing at full stretch, says Emma Ritch, an analyst at business-information firm Cleantech Group in San Francisco, California. Among the companies that have raised hundreds of millions of dollars this year to expand LED facilities are Cree, based in Durham, North Carolina; Seoul Semiconductor, which supplies firms such as Samsung and LG with LEDs for televisions; and Epistar in Hsinchu, Taiwan.

Market forecaster Strategies Unlimited in Mountain View, California, says that the market for high-brightness LEDs will almost triple in five years (see graphic). Vrinda Bhandarkar, an analyst there, says demand for LEDs in mobile phones has peaked, but future hot areas include backlights for better-contrast, thinner computer and television displays. General lighting will follow: LEDs are now bright enough to illuminate streets, but still costly (see *Nature* **459,** 312–314; 2009). Capacity may outpace demand by 2011, says Cleantech, when prices could crash and the herd of more than 500 LED-makers could thin out.

GLOWING PROSPECTS FOR LEDS

A growing market is predicted for high-brightness light-emitting diodes.



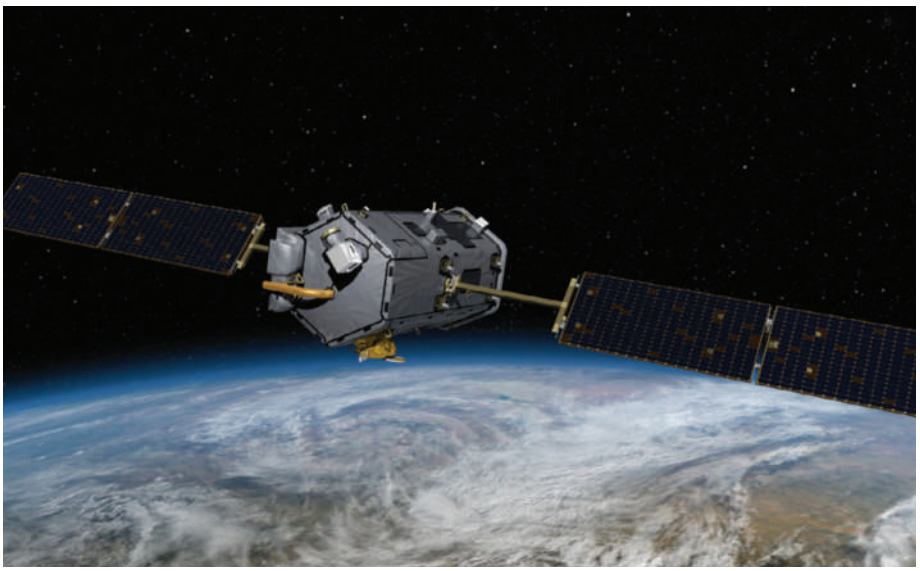

SOURCE: STRATEGIES UNLIMITED

NEWS

# Budget win for climate probe

NASA gets cash to replace a failed carbon-emissions observatory, but concerns remain over future funding.

NASA/JPL



The replacement Orbiting Carbon Observatory will monitor Earth's carbon sinks and sources.

The US Congress is ratcheting up demands for NASA to launch Earth-monitoring satellites that could help to verify the emissions targets currently being debated in Copenhagen.

In a US$447-billion spending bill approved on 13 December (see table), lawmakers told NASA to spend $50 million in fiscal year 2010 on a replacement for the Orbiting Carbon Observatory (OCO), which crashed into the ocean near Antarctica in February after a rocket failure. "It looks like there is a future here," says David Crisp, the mission's principal investigator at the Jet Propulsion Laboratory in Pasadena, California. But by adding the OCO to NASA's already long list of Earth-science missions — and with no promise of future funding — some Earth scientists worry that Congress is asking the agency to do too much. Berrien Moore, director of Climate Central, a think tank in Princeton, New Jersey, says that he was both "pleased and worried" by the OCO funding because of the additional burden on the mission programme.

By measuring levels of atmospheric carbon dioxide, the OCO could provide baseline emissions data and act as a proof-of-concept that carbon sources and sinks can be monitored from space. The observatory would measure $CO_2$ changes to a precision of 1 part per million at a resolution of about 3 square kilometres — nearly 30 times that of the Japanese Greenhouse gases Observing Satellite (GOSAT, also known as IBUKI), which launched in January. Replacing the OCO will cost about the same as the original $280-million mission, says Crisp; if funding continues to be granted, the observatory could be launched as early as 2013.

That would require a much bigger budget for fiscal year 2011, but because NASA is one of several science agencies not included in a targeted doubling of basic-science funding (see 'Will the budget bubble burst?'), it may well face a flat budget next year. "Or worse," says Moore.

NASA already has a list of 15 other Earth-science missions that were identified in a 'decadal survey' to prioritize missions over the next ten years. In the spending bill, Congress said it was "concerned" about the limited progress of those missions, and gave $15 million to accelerate two that are intended to monitor global climate change. It has also instructed NASA to look at using commercial providers, following the lead of a panel that reviewed the agency's human spaceflight programme and earlier this year called for greater reliance on commercial rocket companies.

But Moore doesn't see the Earth-science missions being profitable enough for commercial companies to be interested in running them. In lieu of a surprise windfall in February's 2011 budget proposals, he says, NASA might need to delay the missions further: "We may have to rename the decadal programme the centennial." ■

**Eric Hand**

## Will the budget bubble burst?

US science agencies saw modest budget rises for fiscal year 2010 after Congress approved a spending bill on 13 December. However, researchers are bracing themselves for tighter budgets in 2011 as the government clamps down on its deficit spending.

The National Institutes of Health received a 2.3% rise in funds for 2010, although some observers say that is because the agency is still digesting a $10.4-billion one-off infusion from the 2009 economic stimulus package.

NASA received a 5.3% rise, but funds for its $4.5-billion science directorate were shaved down by $34 million compared with last year. Despite small budgetary windfalls for specific climate-monitoring satellites (see main story), Earth scientists at NASA are worried about the long-term demands on their budget.

Basic-science agencies — the National Science Foundation, the National Institute of Standards and Technology and the energy department's Office of Science — are still on course to double their budgets within a decade, as dictated by the 2007 America COMPETES Act.

However, as the economy recovers and the government shifts its focus from stimulus spending to deficit cutting, some agencies are planning for flat or falling budgets in fiscal year 2011. "It's going to be a tough year," says Patrick Clemins, director of the research and development budget and policy programme at the American Association for the Advancement of Science in Washington DC. **E.H.**

**BUDGET GROWTH FOR 2010**

| Agency | Budget (per cent increase on 2009) |
|---|---|
| National Institutes of Health | $31 billion (2.3%) |
| NASA | $18.7 billion (5.3%) |
| Department of Energy, Office of Science | $4.9 billion (2.7%) |
| National Science Foundation | $6.9 billion (6.7%) |
| National Institute of Standards and Technology | $857 million (4.6%) |

SOURCE: US HOUSE COMMITTEE ON APPROPRIATIONS

R. MCVAY/GETTY

**TISSUE PROFILING**
NMR technology may help surgeons to make the kindest cut.
go.nature.com/QM4s6p

# Royal Institution faces cash crisis

The Royal Institution of Great Britain, once home to historical figures such as Michael Faraday and Lawrence Bragg, has survived since 1799 and is the world's oldest scientific research organization. But it now faces a financial crisis that could bring its 200-year reign to an end.

The institute offers a rare blend of research and outreach, says Richard Catlow, who headed its Davy Faraday Research Laboratory from 1998 to 2007. In the United Kingdom, it is well known for its annual Christmas lectures, a series of high-profile lectures aimed at the general public that are televised nationwide.

But it depends on fundraising and membership for money, and has faced financial difficulties in the past. In 2004 the institution ran up a deficit of £400,000 (US$650,000), according to a 2005 financial statement filed with the Charities Commission, which regulates charities in England and Wales. In 2006, its director, Susan Greenfield, a University of Oxford neuroscientist known for her high media profile, began a £22-million refurbishment of its headquarters in central London, intended to make it a more attractive venue to hire out for conferences and public events. To help pay for the work, the research staff was cut from 60 to just 15, drawing criticism from some scientists (see *Nature* 453, 568–569; 2008).

The project also ran behind schedule and over budget. Fundraising was hampered by the recession, and the institution was forced to dip into its endowment and other 'restricted funds'. By September 2008, it had spent £3.2 million designated for other activities, including the Christmas lecture programme, according to the latest financial statement to the Charities Commission, dated 29 September 2009.

**"Discussions about the role of the director of the Royal Institution are currently taking place."**

Last week, *The Guardian* newspaper reported that Greenfield was being asked to take a pay cut — and reduce the scope of her role — to help make up for the shortfall. The institution declined to comment to *Nature*, saying only that "discussions about the role of the director of the Royal Institution are currently taking place between the board of trustees and the current director".

Chris Rofe, a former administrator at London's Science Museum and the Millennium Dome, was brought in this April to oversee fundraising and financial accountability. The Charities Commission audit acknowledged that a plan was in place to see the organization through to late 2011 and gradually repay the money spent from restricted funds. But it added: "By their very nature, there is a significant uncertainty as to whether these projections will be achieved." ■

Geoff Brumfiel

## SNAPSHOT

## China celebrates panda genome

**With just 1,600 giant pandas estimated to remain in the wild, Chinese scientists have led the task of immortalizing the charismatic critter's 2.25 billion base pairs of DNA, reporting their findings online in *Nature* last week. Although it is unlikely to have a significant effect on conservation, the work is a proof-of-principle for next-generation sequencing technologies, and allows China to trumpet work involving a national animal. Indeed, one tactic for researchers hoping to win funding may be to sequence similarly patriotic symbols. "Australia has the most interesting animals in the world," says Jenny Graves, a geneticist at the Australian National University in Canberra and deputy director of the Australian Research Council's Centre for Kangaroo Genomics, who analysed sequences from the first marsupial (a South American opossum, ironically) and the duck-billed platypus. Graves says that such efforts are not just gimmicks; the kangaroo genomics project has helped researchers to work out that the *SRY* gene determines sex in humans and other mammals (J. W. Foster *et al. Nature* 359, 531–533; 1992). Other patriotic sequencing projects are detailed in the table.**

Brendan Borrell

J. C. MUNOZ/NATUREPL.COM

| Country | Organism | Status |
|---|---|---|
| China | Giant panda | Draft assembly in 2009 |
| Australia | Tammar wallaby | Whole-genome map in 2008 |
| United States (Hawaii) | Transgenic papaya | Draft assembly in 2008 |
| France and Italy | Wine grape (Pinot Noir strain) | Draft assembly in 2007 |
| China and United States | Rice | Draft assembly in 2002 |
| Sweden | Norway (European) spruce | Recently announced |

# Satellites beam in biomass estimates

Additional detail could help bring woodland into a future climate treaty.

Whatever agreement emerges from the climate meeting in Copenhagen, many expect that it will include a mechanism allowing rich nations to offset their emissions by paying poorer countries to protect their forests — and the carbon they contain. But just how much carbon is at stake? Researchers at the meeting have given their best answer yet: the first satellite-based estimates of the biomass contained in the world's tropical forests.

Current biomass estimates for the tropics are based on data gathered by the Food and Agriculture Organization of the United Nations (FAO), and their quality varies greatly from country to country. As a result, baseline figures for biomass are some of the biggest uncertainties in calculating emissions from deforestation and forest degradation, recently estimated to be around 15% of global carbon emissions (G. R. van der Werf *et al. Nature Geosci.* **2,** 737–738; 2009).

The latest assessments, presented at Copenhagen, harness data from multiple satellites as well as thousands of ground plots, and should help governments and other scientists to estimate how much carbon is locked within trees, vegetation and soils on a given patch of land — rather than relying on rough averages that are calculated across a forest.

Sassan Saatchi, an environmental scientist at NASA's Jet Propulsion Laboratory in Pasadena, California, worked on one study with researchers at the carbon consulting firm Winrock International in Arlington, Virginia. He says that their preliminary calculations (see map) accord well with previous estimates. South America comes in with about 145 gigatonnes of carbon in vegetation and soils, about 26% higher than what has been reported by the Intergovernmental Panel on Climate Change (IPCC). The figures for Africa (51 gigatonnes) and south Asia (46 gigatonnes) are about the same as the IPCC figures.

## A question of scale

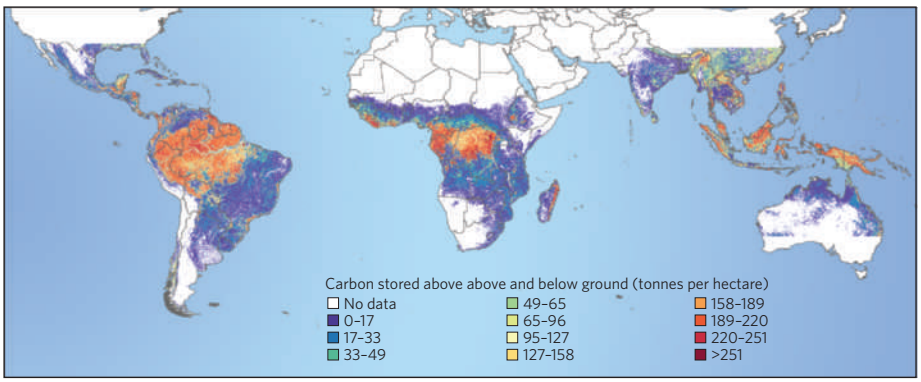
But Saatchi says that the study provides additional information about biomass levels at regional and national levels. "You cannot really nail down this problem unless you have the distribution," he says. "You need to know how biomass is distributed and how it's changing over time, almost everywhere, with some resolution and accuracy."

Funded in part by the World Bank, the work provides a snapshot at 1-kilometre resolution of tropical forests as they were in 2000, when most of the satellite data were collected, as well as more recent deforestation trends.

Also at Copenhagen, researchers at the Woods Hole Research Center in Massachusetts presented another pan-tropical biomass assessment, which had a resolution of 500 metres. Like the Winrock study, it includes spectral data from NASA satellites as well as laser measurements of forest canopy height from an instrument on NASA's Ice, Cloud, and land Elevation Satellite (ICESat) that was designed to study polar ice caps. The two teams have yet to compare results.

Richard Houghton, a biomass expert at Woods Hole, says that it is good news that mul-

THE WORLD'S CARBON STORES




SOURCE: WINROCK INTERNATIONAL

# UK research funding proposal is 'irresponsible'

**Efforts to judge science on its practical returns often raise hackles, and Britain's latest plan is no exception. Some of the nation's leading universities have condemned a scheme that would assess the economic and social benefits of research to help determine who wins a large fraction of university funding, and more than 12,000 academics have signed a petition opposing the plan. The public consultation on the proposals closed this week.**

**Concerns first arose in September, when the Higher Education Funding Council for England (HEFCE) launched its proposals for the Research Excellence Framework (REF), the successor to the Research Assessment Exercise (RAE) used to divide more than £1.5 billion (US$2.4 billion) per year in public funds between universities. Whereas the RAE did not use impact criteria, the REF, due to begin in 2013, will require researchers to submit case studies detailing examples of the societal and economic benefits of their research from the past 10–15 years. These examples would help to determine the fate of almost a third of the funding. The overall aim is to make explicit how much benefit the British taxpayer gets from funding research.**

**Ian Leslie, pro-vice chancellor for research at the University of Cambridge, considers the proposals "neither credible nor responsible". He says that the university recognizes that research institutions need to communicate the impact of the research they undertake. But HEFCE's proposals would turn "first-rate universities into second-rate companies", he says, adding that it is "irresponsible" to apportion so much funding on the basis of the impact of the research.**

**Similar requirements in the United States encourage some researchers to "oversell" the potential impact of their work, says William Schowalter, a chemical engineer from Princeton University, New Jersey, who was an international judge in the final RAE in 2008. This can skew funding towards fields such as nanotechnology that promise more immediate benefits.**

**Peter Knight, deputy rector for**

tiple teams are tackling the big-picture question of tropical forest biomass. "We need a couple of independent estimates just to see how well they match," he says. "Anybody can make a map. If they differ, at least it identifies the areas that need further analysis."

The next step, says Alexander Lotsch, a geographer at the World Bank in Washington DC, is to produce better estimates for carbon emissions from deforestation. He adds that Saatchi's research is still a "work in progress".

Satellites can reliably track deforestation and, increasingly, small-scale logging. In Copenhagen, Greg Asner of the Carnegie Institution of Science in Stanford, California, and Google.org unveiled an online tool that allows tropical countries, beginning in South America, to map deforestation using an automated system to analyse satellite imagery. Asner has also developed a system for assessing biomass at finer resolution, which will be necessary if forests are going to be linked to international carbon markets.

The new pan-tropical biomass maps from Saatchi and Woods Hole won't accomplish that goal, but they can provide scientists and policy-makers with a better understanding of carbon trends. For example, using a similar technique to Saatchi, Asner has found that deforestation in Brazil is moving into higher biomass areas in the interior of the forest. That suggests that emissions will probably rise over time on a per-hectare basis, offsetting some of the reductions in deforestation that Brazil aims to achieve in the coming decade (see *Nature* doi:10.1038/news.2009.752; 2009). ■

**Jeff Tollefson**

**See also www.nature.com/roadtocopenhagen**

**research at Imperial College London, says that he wants to delay the REF by a year to incorporate the findings of planned pilot trials. He recommends that just 15–20% of the audit be devoted to impact assessment, whereas the University and College Union, a trade union for academics, would like the impact component to be removed.**

**David Price, vice-provost for research at University College London, says that any assessment of impact should include benefit to the academic community, and not just the economy and society as currently proposed, to ensure that fields such as mathematics and social sciences are not disadvantaged.**

**HEFCE will publish a summary of the responses to the consultation, and its plans for the REF, in spring 2010.** ■

Natasha Gilbert

PUNCHSTOCK

**COPENHAGEN LIVE**
*Nature*'s reporters blog from the United Nations climate conference.
**go.nature.com/OEsfAa**

# Hope for Japan's key projects

**When Japan's government changed hands in September for the first time in five decades, many Japanese people hoped that the newly powerful Democratic Party of Japan would revitalize their country. But the new government has since sent scientists on an emotional rollercoaster. In recent weeks, two cabinet-level bodies, both chaired by Prime Minister Yukio Hatoyama, have recommended drastically different financial futures for major scientific projects.**

**One set of proposals, from the Government Revitalization Unit (GRU) that was set up in September to trim bureaucratic fat, recommends deep cuts for many key projects. These include a proposed next-generation supercomputer, the SPring-8 synchrotron in Harima, and Earth-science research. Scientists protested against those cuts (see *Nature* 462, 557; 2009). But last week came news of a separate recommendation from the Council for Science and Technology Policy (CSTP), Japan's highest science-policy-making body, proposing continued support for those projects and many others.**

**For instance, the SPring-8 synchrotron and the Global Center of Excellence programmes — meant to strengthen doctoral research programmes — had each been headed for cuts of one-third or more, but the CSTP says they should be "prioritized and given the necessary resources". The next-generation supercomputer, which could have faced outright termination, should also be supported, it says.**

**In Japan, where government decisions are usually made in bureaucratic back rooms and handed out as a harmonious consensus, the apparent contradiction is baffling researchers. "The decision-making process is unclear," says Tadashi Watanabe, project leader for the supercomputer. "It is very unsettling."**

**Final budget decisions will be made later this month, but the prime minister has called the CSTP proposals "valuable opinions", and said that he would "work to ensure they were reflected in the final budget". Many think Hatoyama could be leaning towards accepting the CSTP's recommendations, perhaps because of the outcry over the proposed cuts. "It's too early to tell, but you can safely say that the top leadership did recognize the problem," says Atsushi Sunami, a science-policy expert at the National Graduate Institute for Policy Studies in Tokyo.**

S. KAMBAYASHI/AP

**Scientists are hoping that Japan's prime minister Yukio Hatoyama will avert proposed funding cuts.**

**The GRU was earlier criticized by scientists for recommending cuts without obtaining sufficient external input; projects were usually explained to the decision-making committee in a one-hour session by a bureaucrat. According to the Japanese media last week, the GRU plans to reopen debate on the supercomputer project in a public forum that will involve many scientists.**

**For researchers, the near demise of beloved projects has been a wake-up call to the need to justify them to the public as well as to bureaucrats and external evaluators. Last week, leaders of the supercomputer project posted on their website a list of frequently asked questions on the project's significance, including: "What is great about the supercomputer?"**

**Watanabe says the team doesn't yet have a long-term strategy for engaging the public, but he wants to emphasize that the supercomputer would have a major role in Japan's main research fields, including nanoscience, life science and environmental science. "We have to get that point across," he says.**

**Sunami agrees that scientists in Japan can't take public support for granted. "Even if the budget for the supercomputer and SPring-8 are saved at a smaller scale," he says, "they have to engage with the public more."** ■

David Cyranoski

# Modellers claim wars are predictable

Insurgent attacks follow a universal pattern of timing and casualties.

Seemingly random attacks and a shadowy, mysterious enemy are the hallmarks of insurgent wars, such as those being fought in Afghanistan and Iraq. Many social scientists, as well as the military, hold that, like conventional civil wars, these conflicts can't be understood without considering local factors such as geography and politics. But a mathematical model published today in *Nature* (see page 911) suggests that insurgencies have a common underlying pattern that may allow the timing of attacks and the number of casualties to be predicted.

"We found that the way in which humans do insurgent wars — that is, the number of casualties and the timing of events — is universal," says team leader Neil Johnson, a physicist at the University of Miami in Florida. "This changes the way we think insurgency works."

Johnson and his colleagues argue that the pattern arises because insurgent wars lack a coherent command network and operate more as a "soup of groups", in which cells form and disband when they sense danger, then reform in different sizes and composition. The timing of attacks, the authors say, is driven by competition between insurgent groups for media attention.

Johnson, who has presented preliminary versions of the work to the US military, says that the findings allow a glimpse into the heart of insurgency behaviour. "We can get a sense of what is going on and what might happen if we intervened in certain ways," he says. He is now working to predict how the insurgency in Afghanistan might respond to the influx of foreign troops recently announced by US President Barack Obama.

REUTERS

**Could a model help to predict the number of casualties in conflicts such as that in Afghanistan?**

## Power law

The researchers collected data on the timing of attacks and number of casualties from more than 54,000 events across nine insurgent wars, including those fought in Iraq between 2003 and 2008 and in Sierra Leone between 1994 and 2003. By plotting the distribution of the frequency and size of events, the team found that insurgent wars follow an approximate power law, in which the frequency of attacks decreases with increasing attack size to the power of 2.5. That means that for any insurgent war, an attack with 10 casualties is 316 times more likely to occur than one with 100 casualties (316 is 10 to the power of 2.5).

"This is surprising because these wars are all fought in different terrains and under different circumstances," says Johnson. "It shows that there is something going on in the way these wars are fought that is common to all."

To explain what was driving this common pattern, the researchers created a mathematical model that assumes that insurgent groups form and fragment when they sense danger, and strike in well-timed bursts to maximize their media exposure. The model gave results that resembled the power-law distribution of actual attacks.

"They show a nice agreement between the data and their model, which is an important first step," says Aaron Clauset, who researches the mathematics of conflict at the Santa Fe Institute in New Mexico. But he and others question the model's assumptions, such as the number of insurgents in the conflict remaining roughly fixed over time. Clauset says that this idea does not match with other findings.

The model also assumes that insurgent groups can freely break up then re-form. But Roy Lindelauf, who models terrorist networks at the Netherlands Defence Academy in Breda, notes that some insurgents in Iraq are battling each other as well as the coalition forces, and would therefore not merge into a single group.

Lars-Erik Cederman, a researcher in international conflict at the Swiss Federal Institute of Technology in Zurich, adds that the model "has the potential to improve knowledge about warfare". But, he says, the authors "go too far in claiming they have found a universal underlying pattern" because their work includes only nine wars. Cederman, part of a group that regards insurgency as similar to general warfare, also says that although terrorist attacks can be driven by competition for media attention, it remains far from clear whether insurgencies have the same motive.

"In human social systems, it is usually difficult to nail down what mechanism is behind an observed behavioural pattern," Clauset says. "There are almost always several equally plausible explanations that need to be considered."

Johnson agrees that there could be other explanations for the pattern his group has found. But he says, "We have looked for many years for a model, and this is the only one we have found that explains the data." ■

**Natasha Gilbert**

DECO/ALAMY

**STEM-CELL INDUCTION MADE SIMPLER**
Inserting genes at just one location induces pluripotency.
**go.nature.com/nnItjD**

# Consent issue dogs stem-cell approval

The US expansion of federal funding for human embryonic stem-cell research is being hampered by details in consent forms.

Earlier this month, researchers celebrated the government's approval of funding for a broad variety of work on 13 stem-cell lines — the first approved under the policy announced by US President Barack Obama in March.

But on 14 December, Francis Collins, director of the National Institutes of Health (NIH) in Bethesda, Maryland, decided to respect the unanimous opinion of his standing advisory committee and restrict NIH funding for an additional 27 lines to purposes outlined in the associated consent form. He has as yet declined to discuss his reasoning.

The forms signed by the donating couples stipulated that cells derived from the embryos "will be used to study the embryonic development of endoderm with a focus on pancreatic formation" with the long-term goal of diabetes treatment. Collins's verdict could set the tone for decisions involving some of the 80 other lines awaiting approval at the NIH — and another 242 lines that scientists are preparing to submit.

"This is going to be a recurring issue for the NIH: it is going to look at consent forms that either include restrictive language, or failed to include enough information about the broad array of research possibilities," says Robert Streiffer, a bioethicist at the University of Wisconsin-Madison. "Either of those situations is a problem from the perspective of informed consent."

**"We have to take very literally what is written in that informed consent."**

The cell lines at issue were derived starting almost a decade ago by Doug Melton at Harvard University in Cambridge, Massachusetts. In 2005, responding to a request from Melton, Harvard's institutional review board freed up the lines to be used for "any legitimate scientific purpose". The cell lines had been anonymized, and the donors couldn't be traced for new consent.

But Mary Beckerle, a cancer biologist at the University of Utah in Salt Lake City, told Collins at a 4 December meeting of his advisory committee that "we have to take very literally what is written in that informed consent".

George Daley, a stem-cell researcher at Children's Hospital Boston in Massachusetts, says that despite the restrictions imposed on the Melton lines, federally funded researchers will get the cells they need. "Part of the advantage of having a policy that will have hundreds of new lines is that everybody should be able to find lines that are suitable for their own work," he says. ■

Meredith Wadman

# French research wins huge cash boost

President Sarkozy uses 'big loan' to push his reform agenda.

Universities in France are set to receive an €11-billion (US$16-billion) windfall from a government initiative intended to create an 'Ivy League' of research centres. The cash could help to reverse the decades-long neglect of the country's university system, although the bulk of the funding is likely to be channelled to just a few institutions. This would break a long-standing taboo in France, where the 83 previously centralized state universities have long been considered equal, at least in terms of government funding and researchers' pay scales.

The funding is part of a €35-billion package — the *grand emprunt*, or 'big loan' — announced by President Nicolas Sarkozy on 14 December, intended to boost the country's long-term competitiveness. Borrowed mainly from international financial markets, the investment represents a hefty 1.8% of France's annual gross domestic product, roughly equivalent in scale to the $53-billion science stimulus package announced earlier this year by the United States (see *Nature* **461,** 856–857; 2009).

Sarkozy's spending priorities (see graphic) broadly follow recommendations made on 17 November by a blue-ribbon panel of researchers, industrialists and economists, and chaired by two former prime ministers, conservative Alain Juppé and socialist Michel Rocard. Panelist Edouard Bard, an oceanographer at the Paul-Cézanne University in Aix-Marseille, says he is pleased that Sarkozy heard the panel's argument that increasing the international competiveness of France's universities was the *grand emprunt*'s top priority. The funding should give universities the freedom and funding to attract talent from around the world by offering more competitive salaries and generous lab funding, he says.

## Pillars of excellence

The stimulus funds will be used to create five to ten campuses of research excellence, each of which will receive an endowment of up to €1 billion. The scheme is partly modelled on Germany's multibillion-euro Excellence Initiative, says Bard (see *Nature* doi:10.1038/nature08269; 2009).

> "Big is not always beautiful. It will totally ignore the pockets of excellence in smaller universities."

A further €1 billion was awarded, on top of €850 million allocated last year, to a project to transform the Saclay plateau just outside Paris — one of the highest concentrations of research infrastructure in Europe — into a supercampus by melding many institutions into a single body, fulfilling one of Sarkozy's campaign pledges from the 2007 presidential election (see *Nature* **446,** 847–850; 2007). Other universities will, however, be eligible for other funding, including €1 billion to hire leading researchers or buy equipment to create 'labs of excellence'.

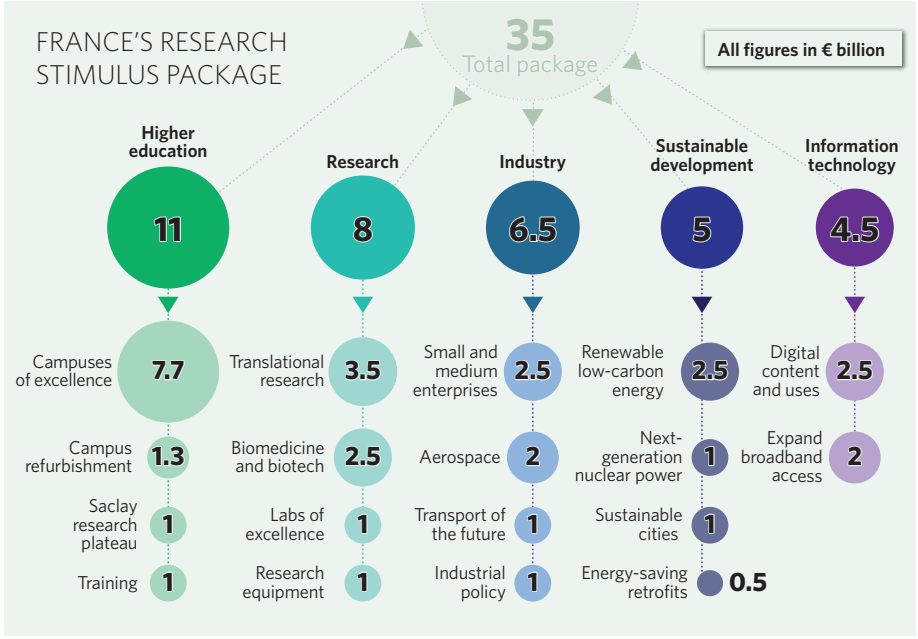
The windfall continues the reforms of a 2007 law that aimed to make universities more competitive by freeing them from central government control and allowing them to manage their own budgets, staff, salaries and buildings. It also reinforces Sarkozy's plan to shift research power away from the national research agencies and towards the universities, and follows the creation in 2007 of a national research council, which awards competitive grants on the basis of peer-reviewed proposals. Most labs had previously operated on a recurrent stream of government funding.

Yet French universities often lack the agencies' expertise in managing large research projects and funds, says Philippe Froguel, a French geneticist working at Imperial College London. Most administrators of French universities, he says, are not in the same league as the heads of major US and UK universities. They will need to develop such expertise quickly if the universities are to spend the new funds well. Arnold Migus, director-general of France's CNRS, the largest national funding agency for basic research in Europe, notes that most university labs are joint CNRS labs, and that the agency is helping universities to get up to speed.

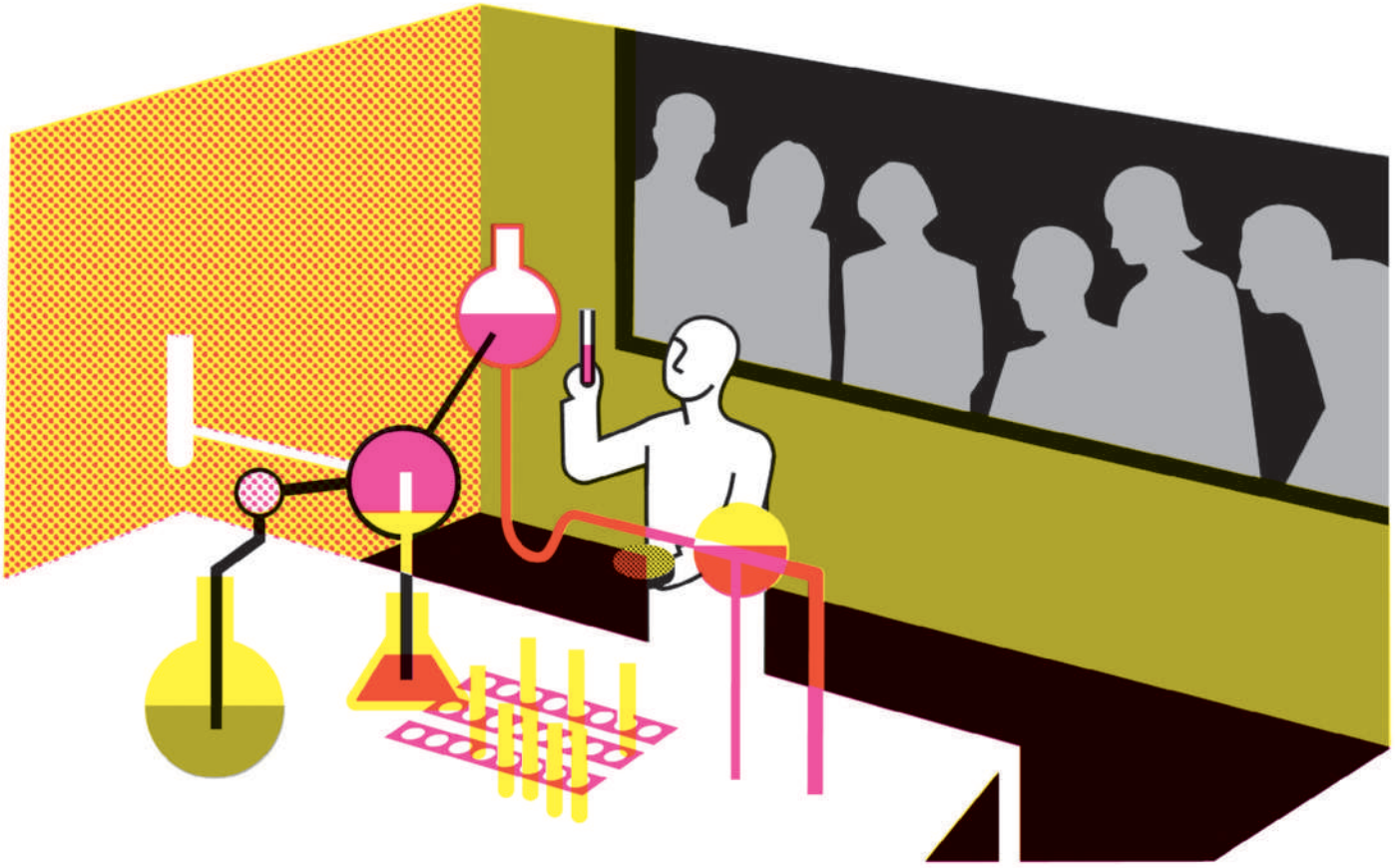
Some French researchers say that the new funds mask cuts in staff and research elsewhere. The "so-called massive investment" is a "mirage", says Bertrand Monthubert, spokesman for higher education and research of the opposition Socialist party, and a mathematician at Paul Sabatier University in Toulouse. He argues that the endowments' annual yields, which are vulnerable to the economic climate, would be similar to the annual budget rises that the universities normally receive.

Froguel applauds concentrating the funding on fewer players, but fears the money will flow only to the largest institutions. "Big is not always beautiful," he says. "It will totally ignore the pockets of excellence in smaller universities." He favours rewarding the best departments or research networks wherever they are, much as Britain's Research Assessment Exercise does (see page 834).

**Declan Butler**




SOURCE: PRESIDENT OF THE REPUBLIC



OTTO

# Watching science at work

A network of social scientists in the United Kingdom is seeking better ways to study the work of biologists. But, asks **Colin Macilwain**, can it earn its subjects' trust?

There was something of a chill in the air at Cardiff's City Hall in October, and not just because autumn was arriving. Social scientists, and some life scientists, were gathering there for the annual meeting of the Genomics Network, a programme run by Britain's Economic and Social Research Council (ESRC) to stimulate dialogue between the disciplines. The plenary talks got under way, led by Christine Hauskeller, a philosopher with the network from the University of Exeter, and Martin Evans, winner of the 2007 Nobel Prize in Physiology or Medicine for his development of gene knockout technology. But when the initial call for questions sparked little real discussion, it was clear that dialogue was going to take some stimulating. Then Evans was asked what he thought of his hosts. "They like to say, why are we doing things?" he growled. "We should be asking, why are they doing things?"

> **"Social scientists have become welcome, and indeed essential, partners with the natural sciences."**
> — Brian Wynne

The Genomics Network started its work back in 2002. The previous year, when the British government had said it would earmark an additional £200 million (US$290 million in 2001) to genomics research, the ESRC successfully argued that about £9 million of it should go to the investigation, by social scientists, of genomics issues and the scientists who study them. Seven years after its creation, the network, which supports about 100 researchers at five universities, is one of the largest projects of its kind in the world, and has broadened its interests beyond genomics to embrace synthetic biology and other areas. Two years ago, the ESRC announced a further £18 million in funding for three of its centres after peer review. The money was for a second, and final, five-year term: the ESRC doesn't support permanent centres of excellence, on the grounds that societal challenges are always changing.

The centres have a very broad scope. Cesagen, the Centre for Economic and Social Aspects of Genomics, is the largest of the groups, and is co-hosted by Cardiff University and Lancaster University. The University of Exeter is home to Egenis, which uses philosophy-based approaches to study genomics questions. Innogen, which studies innovation in genomics and the life sciences, is co-located at the University of Edinburgh and the Open University in Milton Keynes. And the Genomics Forum, also at Edinburgh, was set up in 2004 to help coordinate the Genomics Network and push its findings into wider political and public arenas.

## Embedded in the community

Unlike some previous attempts by sociologists to 'study' scientists at work, in these centres the social scientists are organized into multidisciplinary teams — often including lapsed natural scientists, as well as sociologists and philosophers — with funding to do empirical research. And the researchers embed themselves deeply in the community of natural scientists that they are seeking to study.

Three-quarters of the way though the centres' ten-year lives, their track record is mixed. They have provided a stable and conducive environment for social-sciences research, much of their work is undoubtedly original and some of it has made its impact felt in policy circles, influencing debates on the legislation of animal–human hybrid embryos, for example, and on innovation at the Organisation for Economic Co-operation

and Development (OECD) in Paris. "I think that we've really added value by having these centres," says Ruth Chadwick, the director of Cesagen and chair of the ethics committee at the international Human Genome Organisation. "The ESRC made it clear from the outset that it wanted to see interactions with hard scientists. And we've seen increasing openness to such collaboration: in part because the funding agencies demand it."

Well-trained social scientists can play a much bigger part in helping scientists to build bridges with the outside world, says Grahame Bulfield, former head of science and engineering at the University of Edinburgh and an early champion of the two centres there. "We need organizations that will study the interactions between science and society in a scholarly way, and interpret their findings for the public," he says.

But some critics fault the ESRC for failing to provide sufficient direction for the network since its foundation. Its staff case officer, Liz Grassby, is the fourth or fifth official to hold that position since the network was conceived. "The quality of the research output has not been as good as you might have expected," said one senior social scientist from outside the network. "Much of the problem can be laid at the door of the ESRC; for whatever reason, it has high staff turnover and no collective memory. The initiative could have been better managed."

## Chinks in the wall

The relationship between natural and social scientists has, historically, been more fraught than fruitful. Scientists are often prickly about being studied by outsiders such as sociologists or historians of science.

Lately, however, some chinks have appeared in the wall that separates the two realms. Brian Wynne, a sociologist, former physicist and associate director of Cesagen, based at Lancaster, says that he has noticed profound changes over the past decade in the natural sciences' receptiveness to social science. "Social scientists have become welcome, and indeed essential, partners with the natural sciences," he says. "We've become embedded — like the media folks in the Iraq war," he says.

Wynne, for example, has been involved in planning a citizens' science programme at the Natural History Museum in London, which seeks, among other things, to integrate huge banks of data collected by amateur naturalist groups into the museum's study of biodiversity. Scientists such as Johannes Vogel, keeper of botany at the museum, have built on Cesagen's work on public engagement to help draw on this amateur expertise. For example, the detailed observation of river conditions by fly fishermen led UK regulators to revise their criteria for measuring water quality.

> "Scientists don't especially want to know whether their work produces an ethical dilemma."
> — Tony Woods

In a separate but related project, Vogel, Wynne and two Cesagen researchers, sociologist Claire Waterton and anthropologist Rebecca Ellis, have been contributing to the Barcode of Life initiative, which seeks to build tools that will enable biologists to identify species from short stretches of DNA. Here, the social scientists have been mediating between people who have their own genetic methods for species identification (such as public-health officials checking different strains of mosquito) and the larger international project, which needs global standards for genetic bar-coding. As a result, Wynne says, the whole project is embracing standards, such as on what gene segments to use, that better incorporate existing approaches.

It was James Watson who pioneered the large-scale, systemic involvement of non-scientists in the life sciences when, as associate director for human genome research at the US National Institutes of Health in 1988, he casually suggested that about 3% (later 5%) of all Human Genome Project funds — about $10 million per year since then — should go to the investigation of the Ethical, Legal, and Social Issues (ELSI) involved in the Human Genome Project.

The ELSI programme now serves as the dominant global model for this sort of social-science effort. But it also became synonymous with poor relations between the observers and the observed. From the scientists' point of view "there were two main frustrations with ELSI", says Robert Cook-Deegan, director of the Center for Genome Ethics, Law and Policy at Duke University in Durham, North Carolina. "One was its association with what I call 'finger-wagging ethics'," — telling researchers how they ought to conduct their business. "The other was the way that it created a constituency that wanted grant money more than it wanted to go out and help solve real problems."

Ros Rouse, the ESRC programme officer who built the Genomics Network and is now head of policy at Research Councils UK, the umbrella group for the seven UK research councils, says the ESRC wanted the Genomics Network centres to address "a totally different terrain", from the ELSI programme, including original research into the sociology of biology and medicine, and the nurturing of better links between science, the public and policy-makers.

"This is more of a serious attempt to engage with the life sciences than the original ELSI, which was seen as an 'add-on' to the genome project," says John Dupré, a philosopher of biology who runs Egenis. Dupré is particularly interested in the way that scientists continue to work with a 'tree of life', the representation of species' relationships to each other in a branching tree, even though genomic data challenge it. Genome sequencing has shown that bacteria and other prokaryotes have swapped genes so extensively that their evolutionary histories cannot be represented on a conventional phylogenetic tree[1].

Dupré and his colleagues are now working with philosophers, evolutionary biologists and others to develop other means, such as webs or grids, to represent organisms' genetic relationships. Ford Doolittle of Dalhousie University in Halifax, Canada, is one of the biologists most closely involved. "I think that it's been very useful because biologists don't think very much about philosophy," he says. "We impose patterns on nature for philosophical reasons, and then deny that philosophy is important.

EGENIS

John Dupré (front row, centre) and his team at Egenis study the social and philosophical issues of genomics.

Philosophers care about the structure of arguments and it is a very good exercise for scientists to start looking at this too."

Researchers at Innogen have been examining what Joyce Tait, a former chemist and the first director of Innogen, calls the "innovation triangle" of the pharmaceutical industry, the regulators and the consumer. Tait and her colleagues say that the speed and nature of innovation in some sectors, such as pharmaceuticals and agricultural biotechnology, are determined largely by the regulatory apparatus. They argue, in particular, that it is not the severity of a regulatory system, but rather its ability to discriminate in how it treats large and small companies, that has the greatest bearing on rates of innovation.

These findings are being absorbed in high places. The OECD, for example, used an Innogen study[2] about the future of the pharmaceutical industry as the basis for the health component of a report[3], *The Bioeconomy to 2030*, that it published in June. Given the OECD's considerable prestige, the report is likely to exert a strong influence on government approaches to drug regulation around the world.

## Broad reach

The ESRC always wanted the Genomics Network to reach out beyond academia to public and policy circles, and the Genomics Forum was added with this in mind. "If you have a social-science centre, its prime emphasis is going to be on getting its best work published," says Steven Yearley, a sociologist and director of the forum. "What the forum does is to accept this and make sure that we take that work to a broader audience."

Research teams at the network centres, including Jane Calvert at Innogen and Emma Frow at the Genomics Forum, are involved in helping to shape UK participation in the field of synthetic biology — the search for approaches to build novel biological systems and even entire organisms. The concept of artificial life is expected to stir strong public passions, and those studying it have turned to social scientists from the outset. Some argue that if synthetic biology is to get off the ground, the working dynamic between biologists and engineers needs to be examined and improved. Calvert and Frow have been studying some of these questions, and helping scientists and organizations such as the Royal Academy of Engineering in early public consultations on synthetic biology.

**"Philosophers care about the structure of arguments and it is a very good exercise for scientists to start looking at this too."**
— Ford Doolittle

Alistair Elfick, a medical engineer at the University of Edinburgh, is joint leader of a UK-wide synthetic-biology network that is looking at the technical standards that may be needed if approaches to synthetic biology — such as the design of DNA 'biobricks' that can be pieced together like Lego — are to be successfully pursued. He says he appreciates the social scientists' perspective. "Having their insight will be hugely valuable to us," he says, adding that the nascent discipline is serious about working with social scientists on public engagement. "It's a matter of entering into a dialogue with the public about what it wants us to do. We need to have the authority of society, in order to proceed."

But does the social study of science need to have this kind of practical utility? Some social scientists caution that producing work that is useful to policy-makers, scientists or the public — as the Genomics Network has set out to do — is not always consistent with their core scholarly activity, of seeking to better understand how science works. Paul Martin, a medical sociologist at the University of Nottingham who is not part of the network, warns that the desire to make work fit the needs of policy-makers can create conflicts of interest, in which those who are trying to objectively study science and innovation end up being part of the scientific and innovative process. "We want to be engaged, but not reduced to a handmaiden's role for new technologies," he says.

Nik Brown, a sociologist at the University of York who is also outside the network, worries that policy-makers and scientists can expect the wrong things from social scientists. "There's a misperception that our main role is to ease the interaction between scientists and the public," he says. "What we want to do is understand the science, and how it is constructed — which is not a public-understanding-of-science question."

The biggest practical challenge for the network, though, is the interactions between the scientists and social scientists themselves. One senior researcher, who knows the Genomics Network well, said privately that social scientists still have difficulty speaking a language that scientists can relate to. "There's still a huge credibility gap," according to this observer. "Their methodologies just don't align well."

The issues that interest social scientists most are not always the top concern of many research scientists, who are busy with funding, publications and the science itself. "Scientists don't especially want to know whether their work produces an ethical dilemma," says Tony Woods, head of medicine, society and history grants at the London-based Wellcome Trust, the medical-research charity. "Questions about societies' concerns don't always weigh heavily on their minds."

Asked whether the network has succeeded as whole, Jim Stevenson of the University of Southampton, a psychologist and member of the ESRC's strategic research board, equivocates somewhat: "I think now it is working well, but it has taken a while." He says that the centres need to persevere. "We've got to get to a situation where science, and industry, takes this work seriously," he says. "But it's a slow process." ■

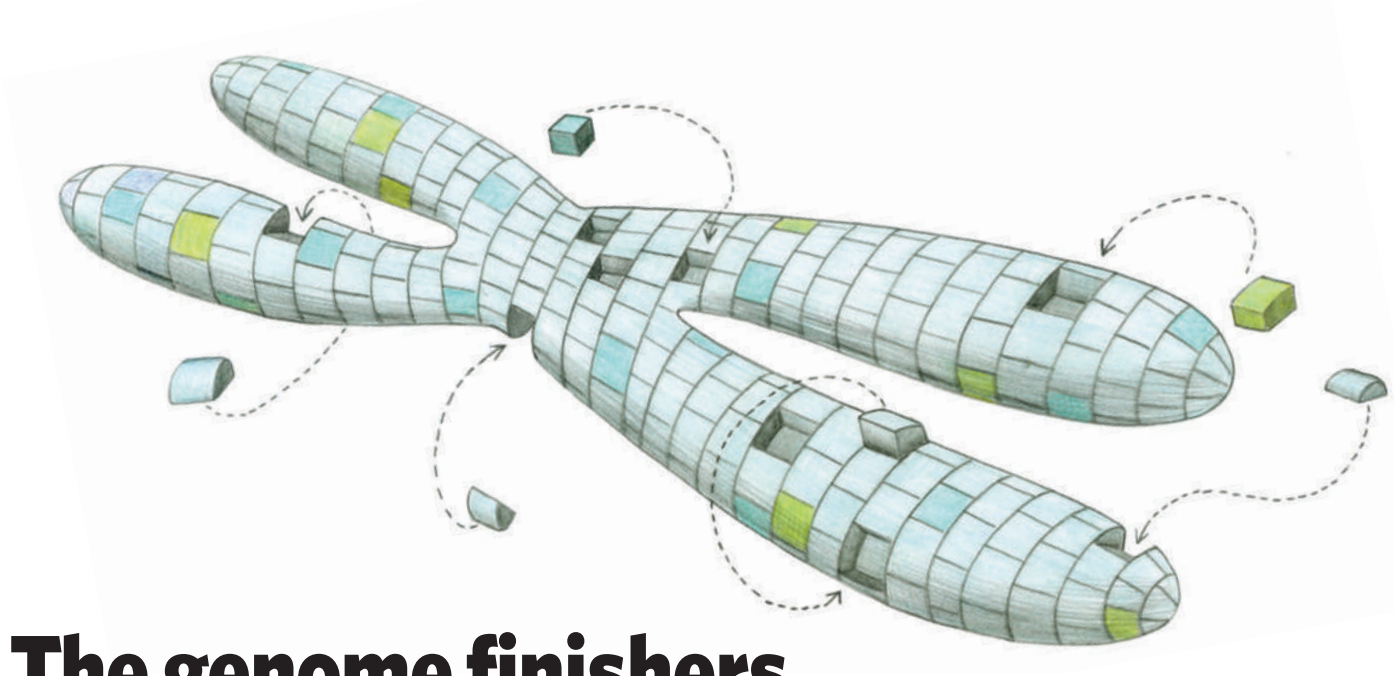
**Colin Macilwain is a writer based in Edinburgh, UK.**


S. OGDEN/SPL

**Social scientists are tackling concerns raised by research to build systems from 'biobricks'.**

1. Bapteste, E. *et al. Biol. Direct* **4,** 34 (2009).
2. Tait, J., Wield, D., Chataway, J. & Bruce. A. (eds) *Health Biotechnology to 2030* (OECD, 2008); available at go.nature.com/YzFDJE
3. *The Bioeconomy to 2030: Designing a Policy Agenda* (OECD, 2009); available at go.nature.com/aQZWFI

**See Editorial, page 825.**



# The genome finishers

Dedicated scientists are working hard to close the gaps, fix the errors and finally complete the human genome sequence. **Elie Dolgin** looks at how close they are.

From her windowless fifth-floor office at the US National Institutes of Health in Bethesda, Maryland, Deanna Church has few distractions from the job that lies before her. On her computer sit 888 open 'tickets', or outstanding problems with the human genome sequence. Although that number fluctuates, it's a not-so-subtle reminder that she and her team at the National Center for Biotechnology Information (NCBI) have a long way to go to finish the job started nearly two decades ago by the Human Genome Project.

This is the same project that an international team of scientists spent close to US$3 billion on to complete. In 2000, the scientists announced, to much fanfare at a White House ceremony, that they had finished the draft sequence of the human genome. They waxed poetic about opening 'evolution's lab notebook' when they published the draft the next year[1]. And they uncorked champagne bottles again in 2003 when the sequence was officially deemed finished[2]. By then, media outlets were reporting the developments with a twinge of fatigue. "This time it is the real thing, scientists promise," *New Scientist* reported. Another year passed before the final analyses were published[3], and two more went by before the paper detailing the last, fully polished chromosome came out in 2006 (ref. 4).

Still, three years later, Church is hunched over her computer, clicking away at her mouse, quietly clearing up the lingering troubles with the iconic sequence. Some of her tickets, submitted by her collaborators and users from around the world, are reports of missing bits. Others describe stretches in which someone thinks the sequence is mistaken. Still others are unique and unexpected challenges, such as complex DNA rearrangements, that could take years to sort out.

"It's a frustration," says Richard Gibbs, director of the Human Genome Sequencing Center at Baylor College of Medicine in Houston, Texas. "It's an extremely high-quality genome. It's the best there is, period. The problem is that a very small percentage of uncertainties still translates into a significant number of problems."

Church and her colleagues are working to build a solid, accurate reference, but their efforts have revealed how slippery that concept can be. The sequence, for instance, does not represent any one person's genome. It is an amalgam of DNA from different people, both male and female. It was put together this way to maintain anonymity for those who contributed the DNA and to ensure that the sequence represented all humanity — "our shared inheritance", as then-head of the project, Francis Collins, said.

**"The work's not sexy. But it's important."** — Deanna Church

But that shared inheritance is hard to capture. The genomes of two individuals look less alike than many had originally assumed. Rather than following a linear path of 3 billion base pairs with a letter changed now and then along the way, human genomes detour into hundreds of vastly different stretches in which, depending on the individual, millions of base pairs can be deleted, inserted, repeated or inverted.

A finished reference genome — if attainable — will therefore look very different from the project's first renditions. That's where Church and her team of finishers come in. They are striving to smooth out the differences and to develop a more dynamic platform that can capture much of humanity's commonalities and uniqueness. Some say it's a wasted effort now that individual human genomes can be sequenced at a fraction of what it cost ten years ago, but most say the reference is invaluable as a bedrock to support the sequencing of future human genomes.

Resolving the problems in the sequence will not win Church many accolades. She won't meet the president or land any papers in high-impact journals as those who "finished" the genome before her did. And once she puts a ticket to rest, there's always another one waiting. "It's not sexy," she says. "But it's important."

## A coalition of the responsible

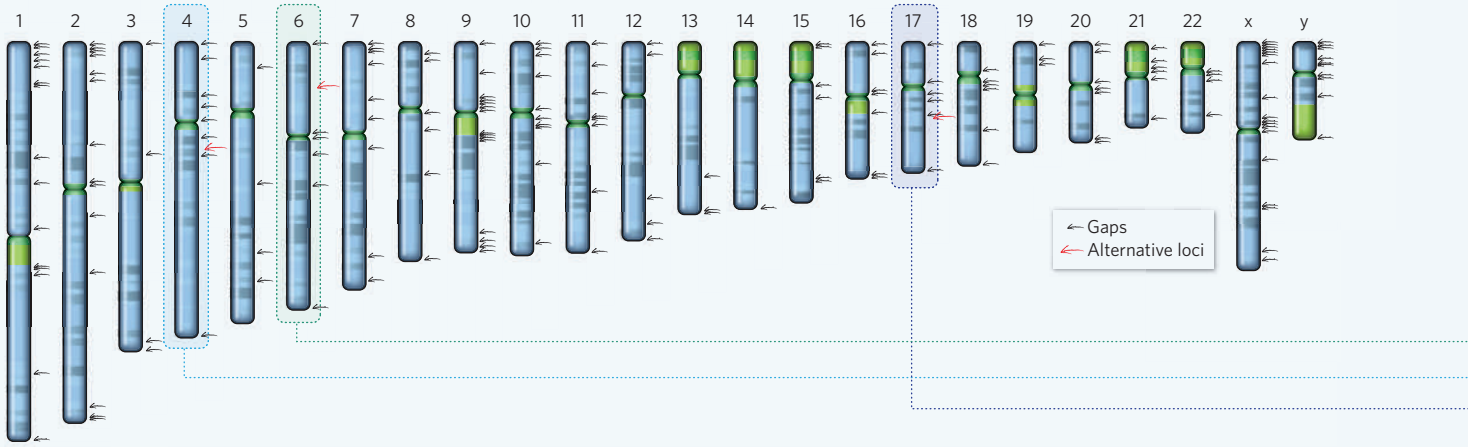
By April 2003, the sequencing had surpassed the international project's technical definition of completion — the sequence contained fewer than 1 error per 10,000 nucleotides and covered 95% of the gene-containing parts of the genome. But there were still errors — around 350 gaps in the sequence — and much of the structural variation was not included.

In 2004, Church and a few dozen researchers met to discuss genomics and structural variation at the Wellcome Trust Sanger Institute in Hinxton, UK. One complaint was echoed repeatedly: there was no easy way to fix or update the genome with new data. In the 1990s, when sequencing was in full

ILLUSTRATIONS BY CHRISTINE BERRIE

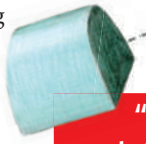
## THE REMAINING GAPS

In March 2009, the Genome Reference Consortium released its first assembly of the human genome that had originally been completed by the Human Genome Project. The sequence closed 25 gaps and added alternative versions at three complex regions (red arrows). The sequence still contains nearly 300 gaps (general locations indicated by black arrows) that range from 700 to 30 million base pairs long, not including parts within the telomeres and centromeres (green) that are intractable to sequencing.




swing, researchers could contact the specific chromosome curators at each of the major sequencing centres involved with the project to report any sequencing slip-ups. But by 2004, few of the centres involved were actively monitoring their slices of the genome and there was little scientific impetus to revisit old work. This posed a problem. "Someone needs to have responsibility for the genome so that if errors are found, improvements can be made," says Adam Felsenfeld, a programme director at the National Human Genome Research Institute (NHGRI) in Bethesda, Maryland.

Together with Ewan Birney of the European Bioinformatics Institute (EBI) in Hinxton, Church appealed to the NHGRI and the Wellcome Trust for funding. It took more than two years of meetings and deliberations, but eventually the NHGRI agreed to set aside up to US$1 million in operating funds from an annual large-scale sequencing award (more than $30 million per year) to Washington University in St Louis, Missouri. Sanger and the Wellcome Trust agreed to a similar amount, and the EBI and NCBI handle the informatics as part of their normal operations. The collaboration, known as the Genome Reference Consortium (GRC), is now the epicentre for genome improvement.



**"Someone needs to have responsibility for the genome."**
—Adam Felsenfeld

To improve the reference, the GRC is concentrating on three main goals: to correct assembly errors; to fill in the genome's remaining gaps; and to produce alternative sequences for regions of the genome with extensive variability.

Researchers have had the first two objectives on their agendas since the Human Genome Project was concluded, and they have been chipping away at them ever since. Some of the regions have been particularly difficult to polish off. For some repetitive stretches, for example, researchers struggled to make multiple copies in bacteria — a necessary part of the sequencing process. But newer methods are now allowing them to fill in these pieces of the genome. Earlier this year, a team led by Chad Nusbaum, co-director of the Genome Sequencing and Analysis programme at the Broad Institute in Cambridge, Massachusetts, used a next-generation sequencing technology that does not require bacteria to amplify the DNA[5]. Nusbaum's team then handed the sequences over to the GRC, which incorporated them into the reference assembly.

The third goal reflects something that has only recently come to light. At first, researchers assumed that genetic variation between people largely consisted of differences in single DNA letters. Now, however, they better understand the extent of structural variations — including deletions, insertions, duplications and inversions. Although some of these variants are involved in heritable disease, they are much more difficult to keep track of than single-letter differences because they often don't map easily to the reference. So instead of representing the genome as a single path of three billion letters, the GRC is introducing alternative paths to reflect its diversity.

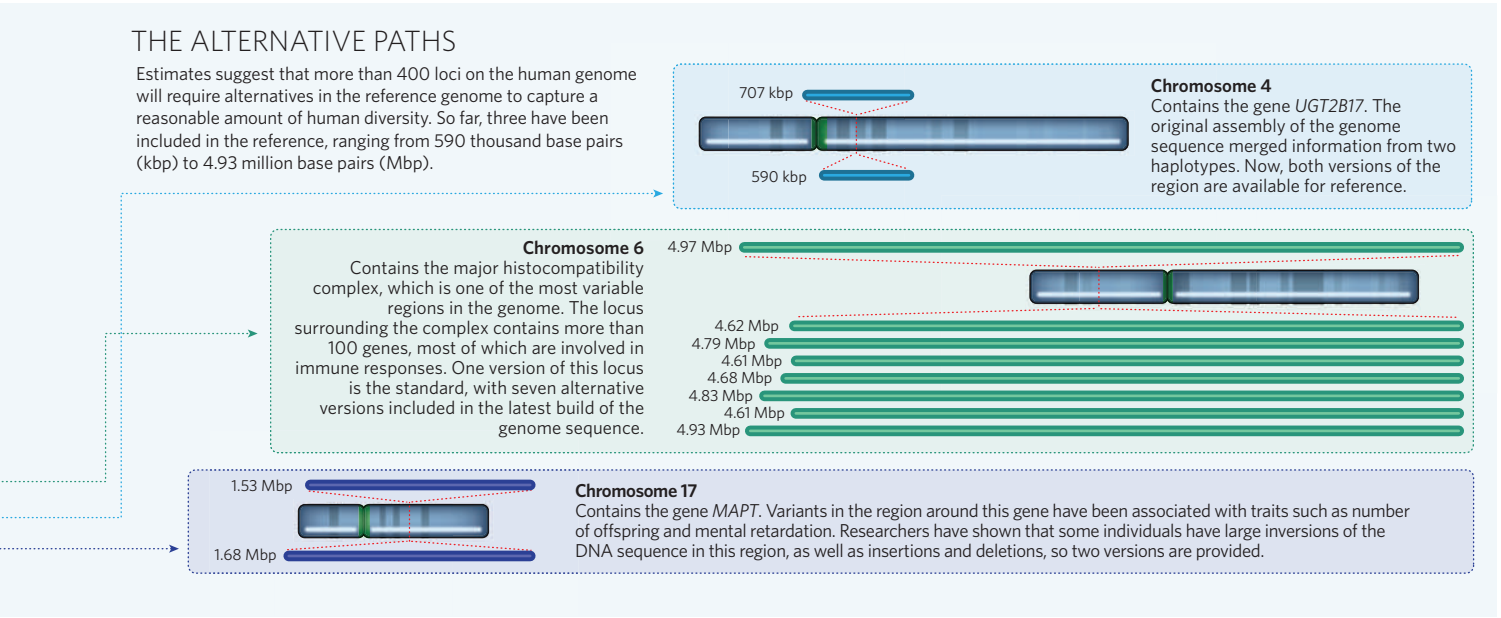
### Twists and turns

One such region is the major histocompatibility complex (MHC) — a 4-million-base-pair stretch of chromosome 6 that contains many immunity-related genes and is recognized as one of the most variable slices of the human genome. The original reference sequence was a hotchpotch of multiple blocks of DNA, called haplotypes, taken from several donors, so the sequence that resulted didn't actually exist in any real person. To create a reference with clearer origins, a team led by Stephan Beck of the University College London Cancer Institute sequenced a single MHC haplotype. They then compared it against seven other common European haplotypes and discovered more than 37,000 single DNA letter differences and around 7,000 structural variations — a level of genetic diversity about an order of magnitude greater than the genomic average[6]. Beck's team's reference has now been swapped into the GRC's default sequence and the other seven haplotypes are included as alternative pathways.

Two other regions also have substitute haplotypes. One sits on chromosome 4 around the gene encoding the UGT2B17 enzyme, which metabolizes steroid hormones and many drugs. The 'finished' reference had misassembled two haplotypes, introducing a false gap. A corrected assembly found that the 'gap' was actually a deletion found only in some people, flanked by large duplications. That section is now included as an alternative pathway in the GRC reference.

The other region, on chromosome 17, encompasses the *MAPT* gene, and provides a case study of the limits of the original reference sequence, which consisted of only one haplotype. An alternative haplotype, a complex inversion of the first that is found in 20% of Europeans, was shown in 2005 to correlate with larger family sizes, suggesting that this second haplotype was under some form of positive selection[7]. But in 2006, Evan Eichler[8], a geneticist at the University of Washington in Seattle, and two other groups[9,10] showed that the inverted region was also prone to frequent spontaneous deletions that led to mental retardation. The inverted haplotype seemed to be both adaptive and the source of debilitating deletions. "You have a yin for the yang here determined by genomic structure," Eichler says. "The question was, 'What's going on?'"

To answer that question, Eichler needed the sequences. He teamed up with Michael Zody, chief technologist of the Genome Biology

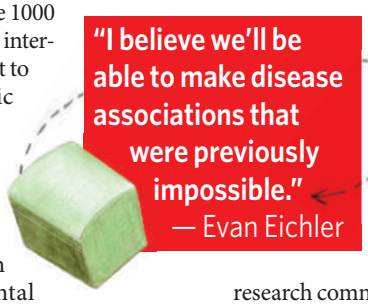



Program at the Broad Institute, to resequence the whole region and showed that the architecture of the inverted haplotype predisposed the sequence to undergo deletions associated with mental retardation[11]. By the time Eichler and Zody published their results, in 2008, the GRC was already in full swing and gearing up to release the next build of the genome. The researchers handed over their sequences to the consortium, and both haplotypes were included in the reference sequence. "The GRC provided a central clearing house for us to go through," Zody says.

Given the clinical relevance of these and other complex regions, providing multiple references is essential to detect the mutations underlying many diseases, says Eichler. "Once we get the alternative structures worked out, I believe we'll be able to make disease associations that were previously impossible," he says. Eichler estimates that around 5% of the genome — corresponding to around 400 specific locations — will need alternative sequences to provide a platform that adequately captures the spectrum of human diversity. These regions include more than 1,000 genes that affect a wide range of physiological processes, including immune responses, drug detoxification and reproductive ability, he says.

## A common task

The GRC's first public offering — a more accurate version of the human genome — rolled out online in March 2009, updated the three parts of the genome with the alternative assemblies, corrected more than 150 alignment problems and closed 25 sequencing gaps. But that still leaves more than 300 gaps. In September, 20 of the GRC's core members gathered in Hinxton for the group's twice-yearly meeting to discuss what steps to take next. While lab workers were clacking away on their keyboards, bioinformaticians were working through one of the consortium's most contentious issues — how to change the reference genome to display only the 'common' gene variants. The GRC's nine-member scientific advisory board, which includes Eichler and Gibbs, has recommended that, wherever possible, the genome should include the common versions of the DNA sequence. But it hasn't defined what 'common' means. Should it be the highest-frequency variant or something that is shared by a reasonable proportion of the population? Should it be calculated across the world's six billion-plus residents or just in a particular ethnic or geographic group? Results emerging from the 1000 Genomes Project, a major international sequencing effort to catalogue human genetic variation in around 2,000 people from across four continents, should inform their decisions.




Some of the GRC members disagree with making such fundamental changes to the reference sequence. "I don't think we should be going through and flipping single bases throughout the genome," says Paul Flicek, who leads the vertebrate genomics group at the EBI. "Informatically, it just doesn't matter. As long as it works, I think it's okay."

Others outside the GRC question whether the entire project is justified. Why bother tinkering with a decade-old reference, asks Lincoln Stein, a bioinformatician at the Ontario Institute for Cancer Research in Toronto, Canada. He calls the effort "more of an abstract exercise than one that's going to have a practical impact". Church, for her part, waves off such criticisms as being from those preoccupied with large-scale genomics. As a detail-oriented person, she knows that the little things count. Individual investigators love their pet genes. That's one reason her queue of tickets is always full. And as genomics increasingly moves to the forefront of personalized medicine, many regions of clinical utility might slip through the cracks. For researchers interested in a particular disease-relevant locus "it doesn't matter that the genome may be 99% complete", she says. "If they're [working with] a region that's incomplete and wrong, they're screwed."

And so the GRC continues with its quiet quest, crossing Ts and sometimes changing them into As, Cs or Gs. Until a reference is no longer needed to assemble the DNA coming from current sequencing technologies, it will continue to document the evolving understanding of human variability in the reference genome. The GRC has also taken on the mouse sequence and will take responsibility for the zebrafish sequence in 2010. Although it may not capture headlines, most in the research community recognize its worth. "The GRC has exactly the right idea," says Jonathan Sebat, a geneticist who studies structural variation at Cold Spring Harbor Laboratory in New York. "It's a no-brainer that someone would be needed to clean up the mess." ■

**Elie Dolgin is assistant news editor for *Nature Medicine* in New York.**

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. *Nature* **422,** 835–847 (2003).
3. International Human Genome Sequencing Consortium *Nature* **431,** 931–945 (2004).
4. Gregory, S. G. *et al. Nature* **441,** 315–321 (2006).
5. Garber, M. *et al. Genome Biol.* **10,** R60 (2009).
6. Horton, R. *et al. Immunogenetics* **60,** 1–18 (2008).
7. Stefannson, H. *et al. Nature Genet.* **37,** 129–137 (2005).
8. Sharp, A. J. *et al. Nature Genet.* **38,** 1038–1042 (2006).
9. Shaw-Smith, C. *et al. Nature Genet.* **38,** 1032–1037 (2006).
10. Koolen, D. A. *et al. Nature Genet.* **38,** 999–1001 (2006).
11. Zody, M. C. *et al. Nature Genet.* **40,** 1076–1083 (2008).

**See Editorial, page 825.**

# Out of service

Decaying infrastructure is an urgent threat that scientists and engineers must help to address, says **Colin Macilwain**.

The trappings of our civilization, from flushing the toilet to posting flip comments on Twitter, rely on a set of critical infrastructures. Many of these — water systems, transport links, electricity grids and generating plants — are ageing severely in developed countries. And the ones that aren't ageing, such as mobile communications and the Internet, are of unknown resilience.

The sudden collapse of a bridge on Interstate 35 in Minneapolis, Minnesota, on 1 August 2007, like the cascade power failure that swept the northeastern United States four years earlier, was a portent of what could come to European and Asian nations if they allow their physical infrastructure to deteriorate.

The US power failure also showed the extent to which rich nations' infrastructure has evolved into a complex web of interdependence, which no one has sought to model properly and for which no authority has overall responsibility. Energy supply, for example, is critical to the operation of all the other infrastructures and is itself dependent on water supply and telecoms.

"The major change over the last 50 years has been the gradual, but ultimately seismic, shift" to an interconnected national infrastructure, where "failure in one part has a direct and damaging knock-on effect in others", noted a scathing report on the topic published earlier this year by the UK Council for Science and Technology (CST), the senior science-advisory body to the British government.

## Under pressure

The problem will be further compounded by global warming. Even before climate change starts stressing existing infrastructure to the limit, the need to cut carbon emissions will transform utilities' priorities. "If you look at most water companies, for example, their biggest bill is electric power. We now have to rethink that," says Paul Jowitt of Heriot-Watt University in Edinburgh, the president of the London-based Institution of Civil Engineers.

Economists, civil engineers and other infrastructure buffs fear that it will take a series of massive failures, akin to the US incidents, for people to sit up and take notice. In the meantime, they hope that scientists and engineers can help to address the problem by demonstrating the hazards posed by the interdependency of networks, and by working with


regulators and government departments on more advanced technological approaches to infrastructure repair and maintenance.

Britain is first in line to confront some aspects of this impending collapse because parts of its sewers, water system and railways date back to the early nineteenth century. The country's problems are compounded by the privatizations of the 1980s, which transferred the national infrastructure from cumbersome but technically competent state bureaucracies to profit-driven entities. None of these firms has a stake in the 50- to 75-year timescales over which infrastructure elements show their worth, and many of them have since jettisoned research and development to save money.

In Britain, according to the Organisation for Economic Co-operation and Development (OECD), investment in water, gas and electricity infrastructure fell from 0.9% of gross domestic product (GDP) in the 1970s to 0.5% in 2000–06. Some places, such as South Korea and Israel, have maintained higher spending levels of three or four times as much. But most, including such dirigiste nations as France, have seen infrastructure spending fall: in the OECD as a whole, it slipped from 1.7% to 0.8% of GDP over the same period.

> **"What we need to be investing in is not consumption but infrastructure."**

All around the world, civil engineers, environmentalists and the business lobby are trying to push infrastructure decay on to the public agenda. And at senior government levels, awareness of the issue is growing. Australia — like Britain, an early pioneer of privatization — passed legislation last year to set up a body called Infrastructure Australia to set priorities and oversee a Aus$20-billion (US$18.2-billion) investment fund. It swung into operation quickly, and this May published a comprehensive set of national priorities.

Also in May, an act proposing a National Infrastructure Development Bank was brought to the US Congress — but it has made no progress. President Barack Obama's stimulus package spread spending very widely, and did little to directly address a US infrastructure spending 'gap' that the American Society of Civil Engineers estimates at a cool US$1.2 trillion over the next five years. "We've made a lot of progress at getting the problem onto the agenda, but not much progress in solving it," says Blaine Leonard, the society's president.

In Britain, Prime Minister Gordon Brown pledged in July to address one of the CST report's central recommendations by establishing a coordinating body, Infrastructure UK. Details of its remit have not been announced.

## Model future

Engineers and scientists, says Leonard, should "prepare to rebuild our crumbling infrastructure with new materials and new technologies, in ways that are more resilient and more sustainable". Better models are also needed to identify critical failure paths in infrastructure, says Brian Collins, chief scientific adviser to the UK transport and business departments and a co-author of the CST report. He says that research groups at the universities of Bristol and Warwick, and at the London School of Economics, have ideas on modelling complex systems that might prove useful.

The UK Engineering and Physical Sciences Research Council, which funds two of these groups, is now pushing national infrastructure as an ideal target for their approaches to modelling. The objective is to identify the most critical of the hundreds of linkages that exist between the different networks.

So where the money will come from to fix these things is anybody's guess. With some exceptions, the institution-building that has taken place has yet to make an impact in spending decisions. Private capital for infrastructure is harder to raise than ever, and most observers fear that public investment plans will collapse after the recession, as they are easier to cut than current expenditure. Dieter Helm, an economist at the University of Oxford, UK, estimated in a September report for the think tank Policy Exchange that Britain alone will need to raise about £500 billion (US$815 billion) for infrastructure over the next ten years.

"You need a crisis first," Helm says. "What you need is for it to visibly start to fall to bits. Ultimately, people will recognize that what we need to be investing in is not consumption but infrastructure." ■

**Colin Macilwain is based in the United Kingdom.**
**e-mail: cfmworldview@gmail.com**
**See go.nature.com/ILx8PC for more columns.**

# CORRESPONDENCE

## Iran's scientists condemn instances of plagiarism

The Iran chapter of the Academy of Sciences for the Developing World, speaking for the country's academic community, deplores the recent cases of alleged plagiarism by Iranian scientists (see *Nature* **462,** 704–705; 2009).

Iran's scientific community is largely free of such unethical behaviour. The calibre of its scientific output is reflected by the substantial growth in recent years in its share of research articles published in high-quality, peer-reviewed international journals.

Several factors account for this improvement in Iran's research output, including sustained and generous government support for science, a swelling of the ranks of young researchers and increasing international collaboration.

The Internet has facilitated communication with our colleagues elsewhere — but the availability of journals on the Internet has also made plagiarism easier. This widely acknowledged problem affects the scientific community worldwide. Iran, sadly, is no exception, and the country's science community is overhauling its practices to counter this scourge.

But circumstances more specific to Iran are also conducive to the spread of plagiarism. Iranian culture places an excessive emphasis on the value of academic credentials, both for advancement in official professions and in social standing. In particular, Iran's political class has an unusual affinity for possessing academic distinctions, as exemplified by the fact that a university degree is a prerequisite for election to parliament. A higher degree is also considered an important qualification for holding other government offices. As a result, the Iranian political class, across the political and ideological spectrum, accounts for a disproportionate share of academic fraud.

**Farhad Ardalan, Hessamaddin Arfaei, Reza Mansouri Sharif University of Technology and the Institute for Research in Fundamental Sciences (IPM), Tehran, Iran**
**Mahdi Balalimood Mashhad University of Medical Sciences, Mashhad, Iran**
**Dariush Farhud, Reza Malekzadeh Medical University of Tehran, Iran**
**Habib Firouzabadi, Keramatollah Izadpanah-Jahromi, Afsaneh Safavi, Shiraz University, Shiraz, Iran**
**Ali Kaveh Iran University of Science and Technology, Tehran, Iran**
**Farrokh Saidi, Abbas Shafiee Shahid Beheshti Medical University, Tehran, Iran**
**Yousef Sobouti Institute for Advanced Studies in Basic Sciences, Zanjan, Iran**

## Opening dialogue between the recent and the long ago

Douglas Erwin's call for palaeontologists to move towards a better understanding of diversity (*Nature* **462,** 282–283; 2009) should be extended. Palaeoecologists should move beyond purely descriptive objectives and towards a better understanding of ecosystem evolution.

The worlds of palaeoecologists and ecologists are very much apart, and dialogue between the two remains muted. (One exception is the Quaternary, in which several key researchers have had a strong ecological background.)

Palaeoecologists working across geological time should, we suggest, familiarize themselves with the paradigms being debated in the ecological literature, and seek ways with which they could be examined in the fossil record. Palaeoecological research has to rely on uniformitarian assumptions using modern ecological analogues. Similarly, ecologists could be more sensitive to the limitations of data collection from the historical record, and could help frame modern studies to complement palaeoecological hypothesis testing.

The way modern and fossil ecosystems are described will necessarily be different, but researchers should strive to see whether the rules described for modern ecosystems hold over geological time. After all, the fossil record provides the only way to study changes in ecosystems over more than centennial timescales, along with samples of non-analogous ('extinct') ecosystems and habitats.

As first steps, palaeoecologists could publish in leading ecological journals and participate in ecological conferences.

**Julien Louys, Laura C. Bishop, David M. Wilkinson Research Centre in Evolutionary Anthropology and Palaeoecology, School of Natural Sciences and Psychology, Liverpool John Moores University, Liverpool L3 3AF, UK**
**e-mail: j.louys@ljmu.ac.uk**

## UK defence group's structure could limit its usefulness

You describe a new UK Ministry of Defence programme, in a News story (*Nature* **462,** 151; 2009), as being modelled on its US counterpart, the JASONs: this is an independent group of scientific advisers with high-level security clearance who have consulted for the US government on technical problems since the cold war. As a member of the JASONs for more than a decade, I note that several organizational aspects are crucially different.

First, the British government intends to finance research in the laboratories of members of their advisory group in order to follow up on ideas that they generate. Second, most members won't have security clearance. Third, members will not be self-selected, but will be appointed by the government. Fourth, the membership will be rotated frequently according to topic.

In my opinion, these differences detract from the usefulness of the UK enterprise. Lack of clearance discourages members from bringing up relevant issues, and will hobble security-related work generally. Controversial conclusions of reports could be dismissed by government officials with the all-too-familiar refrain: "You wouldn't advise that if you knew what we know."

Also, the group stands to forfeit independence because the government will select scientists according to its own criteria, which tend to include political expediency in addition to competence. They risk losing credibility within the scientific community for the same reason. When members are funded to pursue projects at their home institutions, it raises conflict-of-interest issues. And rotating the membership will prevent the organization from ever developing a true 'corporate memory'.

As an organization, the JASONs have earned credibility on a wide range of security-related topics over the years. The long-standing working relationships forged among the JASON members rank among its most significant strengths. The lack of direct professional ties to government sponsors fosters impartiality. The collective experience gained from working on so many different kinds of problem, combined with the individual credentials of its scientifically diverse membership, make the JASONs one of the few government advisory groups that can plausibly be called independent.

**Steven M. Block Departments of Biology and Applied Physics, Stanford University, Stanford, California 94305-5020, USA**
**e-mail: sblock@stanford.edu**

Contributions to this page may be sent to correspondence@nature.com. Please see the Guide to Authors at go.nature.com/cMCHno. We also welcome comments at Nautilus (http://blogs.nature.com/nautilus).

OPINION

# Geothermal quake risks must be faced

Discussion needs to be open about how exploitation of Earth's internal heat can produce earthquakes, says **Domenico Giardini**, so that the alternative-energy technology can be properly utilized.

Deep geothermal energy is increasingly being explored as an attractive alternative energy source. Conventional hydrothermal resources, such as hot springs in geothermal areas, have been effectively exploited in the past century, but their distribution and potential for supplying electricity is somewhat limited. Tapping deep geothermal energy offers new prospects.

An enhanced geothermal system (EGS), originally called a 'hot dry rock' system, involves drilling a hole at least 3 kilometres deep into a layer of non-porous rock where temperatures are higher than 100 °C. Fluids are pumped under high pressure into the rock (a process called stimulation), which induces it to fracture, generating micro-earthquakes, thereby increasing its permeability and creating a reservoir for the fluid. The ruptures generate elastic waves that are detectable by sensitive seismic networks. Once a reservoir of permeable rock larger than a cubic kilometre has been formed, additional holes are drilled to extract heat from the rock mass by circulating fluids through the fracture network.

The brute-force approach of EGS is attractively simple. And it has, theoretically, the capacity to generate large amounts of alternative energy by tapping a virtually unlimited source — the heat stored deep inside Earth. An expert panel convened at the Massachusetts Institute of Technology in Cambridge in 2006 estimated that EGS could provide up to 100,000 megawatts of electricity in the United States by 2050, or about 10% of the current national capacity — a very large proportion for an alternative energy source. In October, the United States announced that up to US$132.9 million from the recovery act would be directed at EGS demonstration projects, and big names including Google have invested in the technology.

The drawback is that such enhanced geothermal systems can induce earthquakes. The initial stimulation creates micro-earthquakes that might be felt at the surface or even produce damage. And the pressurized water forced into the rock could interact with existing deep faults, generating potentially large quakes. The probability of this happening is not large, but needs to be considered. In addition, geothermal energy is more profitable if it generates electricity and heating at the same time. That means that customers have to be close to the energy source, so it is attractive for operators to develop geothermal-energy sites in urban areas, where earthquakes are more problematic.

J. WILSON/*THE NEW YORK TIMES*/REDUX/EYEVINE

**Enhanced geothermal systems, such as this planned one in California, must undergo quake risk analysis.**

Thousands of deep geothermal sites will have to be developed for geothermal energy to supply a sizeable component of the global energy need. If a significant fraction of these induce seismic action under dense urban areas that is felt or is damaging, this will exceed the natural rate of activity in stable continental areas. Man-made rather than natural earthquakes are already the dominant component of seismicity in mining districts in countries such as Poland and the Czech Republic, but is society across Europe and elsewhere ready to accept this threat in urban areas?

In a recent case in California, a planned EGS site at the Geysers, a geothermal power field about 100 kilometres north of San Francisco, met with public resistance and fell under review by the Department of Energy (even though the company involved had completed an appropriate seismicity review). In September, that project was suspended because of technical difficulties.

For an enhanced geothermal system located near a city or in an area already hit by past large earthquakes, the increased seismic risk requires developing mitigation strategies, such as restricting the pressure or location of pumped fluids. Open and comprehensive information and education needs to be provided to the public and to authorities before, during and after the project. The risks must be openly recognized and assessed, and thought needs to be given to how to insure against damage caused by the projects. Discussion is needed with all stakeholders — including scientists, politicians and the public — to decide what level of risk is acceptable. Otherwise society risks a public backlash that could unnecessarily quash a promising alternative-energy technology.

## The Basel story

One of the first purely commercially oriented EGS projects — the Deep Heat Mining project — was initiated in Basel, Switzerland, in 1996 by the Geopower Basel (GPB) consortium. In my view, what started as a promising green-energy initiative turned into a messy affair. It is a textbook example of how the failure to come to terms fully with the possibility of producing earthquakes in an urban area (by everyone involved — including the public) became in itself the largest risk to the whole

concept of geothermal exploitation. We can learn important lessons from the case, which should serve in securing a long-term future to this promising energy source.

Basel, an industrial centre of Europe's chemical and pharmaceutical industry, borders France and Germany, and more than 700,000 people live in the area. It has a history of earthquakes; in 1356, the city was severely damaged by a magnitude-6.7 quake, the largest ever recorded in central Europe.

Preparing for a commercial EGS project in an industrial zone took several years. In October 2006, the injection well reached its final depth of 5 kilometres, and was ready for the injection of high-pressure fluids into the granite. A monitoring system was installed, with six borehole seismometers installed near the injection well and up to 30 seismic surface stations in the Basel area, and a contingency shutdown plan in case of felt earthquakes. Nevertheless, the Swiss Seismological Service, which had no regulatory power in this case, communicated to GPB and the Basel authorities that the service had not seen what it would consider an adequate seismic risk analysis for the project.

The local authority confirms that GPB had a valid permit, and had met all that permit's conditions. On 2 December 2006, GPB began injecting water into the well with increasing flow rates. As expected, thousands of micro-earthquakes were recorded. Because of the strongly increased seismic activity felt at the surface, injection was stopped on 7 December. A few hours later, a magnitude-3.4 event rattled the local population, causing fear and anger, and receiving international media attention. In a press release on 9 December, GPB announced its regret for the incident, saying the tremors produced by the project were larger than expected. Slight non-structural damage, such as fine cracks in plaster, was claimed by many homeowners and paid by GPB's insurance. The incident also led to a court case against an individual — not GPB — that starts this week.

Since the water injection stopped, seismicity in the area has slowly decayed. Three years later, sporadic seismicity inside the stimulated rock volume is still being detected by the down-hole instruments.

This EGS project has been on hold, awaiting the completion of an independent risk analysis by a consortium of seismologists and engineers, selected by state authorities following an international bid. The study was released on 10 December this year, and public authorities have now decided to suspend the project.

There have been several other forays into enhanced geothermal energy projects in Europe, some of which have been associated with earthquakes. The European Hot Dry Rock geothermal-energy project in nearby Soultz-sous-Forêts, France, has been developed to a depth of 5 kilometres over the past decade. During stimulation, seismicity was generated there with a maximum local magnitude of 2.9. The plant was adapted to reduce the earthquake risk, and is scheduled to begin producing electricity in January 2010. At 2 megawatts, it will be the largest commercial EGS site in operation. Felt earthquakes are also occasionally associated with natural geothermal systems. In Landau, Germany, a 3-kilometre-deep system was constructed in naturally permeable layers, and earthquakes were not expected. However, seismicity was felt a year after the start of energy production, in 2007, and suspended operations for many months. In both of these cases the geothermal exploitation is carried out in more-rural areas without a known history of large earthquakes.

Sweeping up: a geothermal project in Basel, Switzerland, has been suspended.

C. FLIERL/*THE NEW YORK TIMES*/REDUX/EYEVINE

## Realistic approach

The risk of overreaction to the risks inherent in deep geothermal projects is very real. The establishment of an overly harsh regulatory framework would penalize the geothermal industry in comparison to other energy sectors that carry a recognized risk of inducing seismicity, such as gas extraction or coal mining.

> "Society risks a public backlash that could unnecessarily quash a promising alternative energy technology."

From their outset, EGS projects need to be thought of both as pilot projects with scientific unknowns and as commercial ventures with technological and financial risks. Companies need to have allocated enough of their budget to scientific investigations not directly related to the exploitation of heat. Local authorities need to avoid being enticed by the promises of alternative energy, and to remember to ask the right questions. Risk evaluations need to be done before — not after — these projects begin.

Even if the right questions are asked at the right time, the scientific and engineering community is hard pressed to provide a consensus opinion on how seismic hazards can be assessed with confidence and minimized. The empirical data include only a handful of well-monitored EGS experiments; models are consequently poorly constrained. The European Commission has approved the Geothermal Engineering Integrating Mitigation of Induced Seismicity in Reservoirs (GEISER) project to improve the knowledge base and suggest procedures and regulations for the future exploitation of deep geothermal energy. However, many EGS projects are expected to open in the years before the GEISER project produces useful results.

The Basel programme is likely to have a strong effect on the insurance cost of future projects associated with induced seismicity. The damage claims in Basel amounted to more than $9 million, which seems a high toll for a local magnitude-3.4 event (although this is hard to say definitively, because data on small non-structural damages from past earthquakes have never been comprehensively collected). The damage in each building never reached the 10% property level that is normally applied as deductible by home insurance policies. For a natural event, the damage would have been covered by the homeowners, but for a man-made event, the whole cost was picked up by the company's liability insurance. This of course opens a difficult issue. How would we treat a magnitude-5.5 earthquake hitting Basel in, say, 30 years? Could we prove whether it was natural or not? Who would cover the damage?

The public reacts with a vengeance if it perceives that a known problem has been hidden. More than this, earthquakes invariably raise primordial fears. Waking up the sleeping terror that lurks in the deep is the plot of numerous horror movies; here it has an all-too-real meaning.

It is now becoming clear to the public, local authorities, the geothermal industry and regulatory agencies that deep geothermal systems carry a small risk — as do most technologies in the energy sector. Dams can break, nuclear power plants may fail, carbon dioxide released from the oil and gas contributes to global warming, and EGS projects can create damage through induced earthquakes. The open question is whether or not society is able to find ways to balance and accept these risks. A well-informed discussion is needed to find out. ■

**Domenico Giardini** is director of the Swiss Seismological Service, ETH Zurich, Sonneggstrasse 5 , CH-8092 Zurich, Switzerland.
e-mail: giardini@sed.ethz.ch

**See go.nature.com/rk5jgK for further reading.**

BOOKS & ARTS

# A vision of the nanoscale

A collaborative effort between a photographer and a chemist could show scientists how to make the small scale more intuitive, says **Jeremy Baumberg**.

**No Small Matter: Science on the Nanoscale**
by Felice C. Frankel and George M. Whitesides
Belknap Press (Harvard University Press): 2009. 192 pp. $35, £25.95, €31.50

F. FRANKEL, *NO SMALL MATTER*

**This glass apple's curious shadow (part round, part cube) symbolizes the duality of the quantum world.**

On the bookshelves of my childhood, I remember science encyclopaedias that were filled with images of natural and synthesized wonders. No modern shelf should be without titles that chart today's exploration of science's outer limits. But whereas hoards of glossy books have been published on astronomy, for example, the small world has been shyer to emerge. Perhaps this is partly because nanoscience, like postmodern art, seems to demand a lengthy explanation of what we are viewing — but at the expense of an unmediated experience.

Reorienting our eye to the nanoscale is *No Small Matter*. This coffee-table book juxtaposes images and ideas to encapsulate the significance of size and shape. It is the product of a second collaboration between science photographer Felice Frankel and chemist George Whitesides, and follows their project to capture images of pattern formation on surfaces (*On the Surface of Things*; Harvard University Press, 1997). Exploring where art meets science, the authors search for promising paths to make small-scale science more intuitive.

Science–art collaborations frequently end up as one-way processes, with science donating the metaphors to a consumerist arts culture that fashions them into a new end product. Frankel avoids this trap by using images — photographs on a human scale and through microscopy — to frame graphic icons that encapsulate a theme and draw in the viewer. As with much of art, the explanatory power of these pared-down representations relies on the previous experiences of the person who enquires of it, so the collection is a hit-and-miss affair. For instance, my response to their photo of an abacus that introduces the theme of counting in binary — clicking, serried ranks — might be very different from yours.

Partnered with each photograph is a short, blog-like column. The text informs while the image reinforces the text's metaphors, each pair offering a self-contained journey. The book's themes — conveyed in textbook-like sequence by bold numbering — include structural and conceptual underpinnings at the small scale; life and engineering; and nanotechnology's potential risks and prospects. It is a highly personal tour of science that benefits from the clear intuition born from Whitesides' experience. Such thoughtful tactics epitomize the difficulty of making the revelations of nanoscience accessible.

The weaving of an eminent nanoscientist and an inspired photographer creates a sometimes inconsistent cloth. Some images inspire awe, such as the delicate cage of the marine sponge known as Venus's flower basket, or the grasping arrays of polymer rods described as nano-fingers. Others lack depth, such as the ubiquitous candle flame. And yet others don't quite illuminate, such as a depiction of the quantum world through an artificially rendered cubical shadow of a spherical apple. Forcing such visual simplicity leads to metaphors that hang flapping — for example, a full or empty wine glass caricatures the digital nature of a binary bit, but it doesn't help to explain the elegant binary logic of $1 + 1 = 0$. Nonetheless, the authors' engagement is stimulating.

The book does convey the idea that shape and size matter hugely to nanoscale function. It captures well the mystery of how the layering of these simple concepts generates the complexity of life. But the authors' attempts to justify the funding and ambition of nanoscience are awkwardly conflated with societal challenges, such as clean-energy generation. Although there are connections, they don't emerge simply from the preceding material, leaving the reader bewildered.

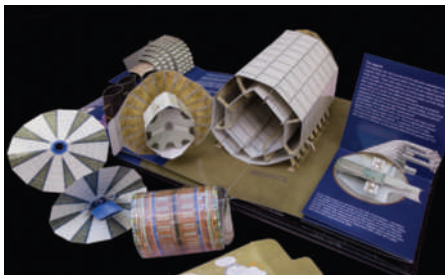
The big-science communities of particle physics, genomics and astronomy are highly organized when disseminating knowledge and lobbying on behalf of their fields. Despite its much greater number of practitioners, the nanoscience crowd has not spread its messages across wider society. Using images such as these to build people's familiarity with nanoscience can provide a visual shortcut that connects emerging stories with a broader message.

Frankel and Whitesides' book adds gravitas and nuance to the popularization of nanotechnology, articulating its interest and vast opportunities. Rather than being stuck on a teenager's bookshelf, *No Small Matter* should lie open on our coffee tables, inviting comment. ■

**Jeremy Baumberg** is professor of nanophotonics and director of the Nano Doctoral Training Centre at the Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK.
e-mail: j.j.baumberg@phy.cam.ac.uk

# Pop-up physics

**A Voyage to the Heart of Matter: The ATLAS Experiment at CERN**
by Emma Saunders with Anton Radevsky
Papadakis: 2009. 8 pp. £20



Detector origami: the pop-up model of ATLAS.

PAPADAKIS PUBLISHER

Finally up and running after a 14-month repair job, the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, is the most complex experiment in the world and took thousands of physicists well over a decade to assemble. So it is perhaps to inspire empathy that an LHC pop-up book asks its readers to spend a few minutes fumbling with pieces of paper that (eventually) fold into a model of the giant ATLAS detector, one of four detectors at the collider.

This feature and others do not necessarily mean that *A Voyage to the Heart of Matter* is aimed at young children; indeed, its author, Emma Saunders, admits that she would not let her three-year-old son flip through it unsupervised. But it is perfectly suited for any adult with even a passing interest in the LHC and a desire to try their hand at detector assembly.

The book manages to pack an incredible amount of information about the collider into just four spreads. The first guides readers through the layout of the collider and its design; the second and third describe the ATLAS detector in detail; and the final spread, the book's most beautiful, shows galaxies in flight around a swirling maelstrom of primordial particles similar to those being hunted by the LHC physicists. In addition to a main pop-up, each page contains four flaps that open to reveal further facts about the collider and its goals.

The book's greatest charm is its obsession with detail. There is no reason one would have to include an 'inner detector' in the ATLAS mock-up; yet physicists would surely object if it was not there. Similarly, the fountain-like recreation of the Big Bang, in addition to being beautiful, is highly accurate. At its base — corresponding to The Beginning — lie unconfined quarks, while the upper layers correspond to the cosmic microwave background, the beginnings of large-scale structure, and the eventual formation of galaxies.

Most important, the pop-up format actually improves the book. Rather than being gimmicky, each pop-out genuinely illuminates the workings of the detector and the interactions of the particles it hopes to find. The book would be the perfect holiday gift for any armchair physicist who wants a little taste of life at the LHC. ■

**Geoff Brumfiel** is a senior reporter for *Nature*.

# Trust puts the self on show

**Identity: Eight Rooms, Nine Lives**
Wellcome Collection, London
Until 6 April 2010

A maze of corridors winds through featureless partition walls, adorned only by mirrors ranging from the exotic (Etruscan artefacts from the collection of Sigmund Freud) and the surreal (a digital 'time-lapse' mirror in which one's past movements materialize belatedly, like a ghost) to the banal (a cheap, plastic shaving mirror once belonging to actor Michael York). The corridors feed into eight rooms, each exploring an aspect of the knotty question of who we are.

*Identity* launches a series of national events organized by UK science-funding body the Wellcome Trust. At this showing, Ken Arnold, head of public programmes at the Wellcome Collection in London, announced that the upcoming tenth anniversary of the draft human genome seemed a timely moment to step back with a mix of medicine, popular culture and high art, designed to get at a "dauntingly broad" question. "We quickly found that philosophy and neuroscience really didn't have answers to some very basic questions about identity," said one of the co-curators, Hugh Aldersey-Williams. "Meanwhile, biology, society and law find themselves increasingly in conflict over these basic issues." Their strategy in tackling the vast subject was to choose eight topics, each introduced by a figurehead personality.

One of these is geneticist Alec Jeffreys. Behind glass are artefacts from his childhood: a well-loved copy of *Biggles Works It Out*; a school report displaying all A grades (including 19.5 out of 20 in science), apart from a B in writing and a C+ in physical training and games ("tries hard"). In another case, the trappings of Jeffreys' scientific life are arrayed like ancient relics: a battered Geiger counter, an X-ray film with scattered black bands; a pivotal *Nature* paper (*Nature* **317**, 818–819; 1985). Among these everyday tangibles, Jeffreys's eureka moment is mapped with remarkable precision: "At 9 a.m. on Monday 15 September 1985, he found what he was looking for," says the placard. "By afternoon he had coined the phrase 'DNA fingerprinting'."

Unlike the other isolated rooms, Jeffreys's is fused with that of Francis Galton, the Victorian polymath who pioneered fingerprinting as a means of identification. Aldersey-Williams confesses to an attempt to draw "cheeky parallels" between the trivialities of the two men. Thus,


Francis Galton and Alec Jeffreys (centre) both pioneered fingerprinting: the first using ink, the latter, DNA.

WELLCOME IMAGES

the childhood memorabilia of Galton includes a worn copy of the *Iliad* and a letter to a relative, in which the four-year-old brags about being able to read "any English book" and recite "all the Latin substantives". Galton's slant on identity was all about phenotype: on display are the composite photographs with which he tried to distil the facial essence of particular groups: criminals, asylum patients, Baptist ministers — even scientists. Further parallels are introduced in the room devoted to physiologist Franz Joseph Gall, which is full of the trappings of phrenology — moulds of crania and masks of faces, including those of Voltaire and Isaac Newton. These are juxtaposed with videos of scans using magnetic resonance imaging, which show areas of the brain lighting up during difficult decision-making or jazz improvisation — the sort of modern phrenology of which Gall might have approved.

Other rooms explore the perceptions of self from a less scientific vantage. One displays diaries, including that of Samuel Pepys and of Clive Wearing, a pianist with anterograde amnesia whose inability to form new memories means that every line reflects perfect immediacy, as if recording the first moment of his life. Visitors are invited to sit down in the actual *Big Brother* chair and mingle with traces of D-list celebrity sweat. There is a room about twins. In it, two identical twins struggle to distinguish themselves as they age, but the series of photos culminates in an adult pose that is unconsciously mirror-image: the result, we are told, of an egg that split early on in development.

The genetics of gender is addressed in a room about April Ashley, one of the first people in Britain to have a complete sex-change operation. A succession of press clippings charts not the transformation of someone from male to female, but rather the relentless battle of someone who always thought she was a woman. This paper trail ends in a retrospective change of gender on Ashley's birth certificate — a reinventing of biological history that might feel vaguely uncomfortable to the average visitor. Yet the recent public outrage on behalf of Caster Semenya, the world-champion middle-distance runner whose gender is still in dispute, shows that many of us are prepared to define sex by means other than strictly genetic.

Co-curator James Peto admits that, at the beginning, he thought covering a topic such as identity was completely overwhelming. "You can only scratch away at little bits and hopefully raise enough questions for people to start finding answers for themselves." ■

**Jennifer Rohn** is a cell biologist at University College London and editor of LabLit.com. Her first novel is *Experimental Heart*.
e-mail: jenny@lablit.com

Mona Hatoum's *Hot Spot* globe.

K. WIGGLESWORTH/AP

# Artistic dispatches on climate

**Earth: Art of a Changing World**
Royal Academy of Arts, London
Until 31 January 2010

Photographs of our blue planet, taken during the 1968 *Apollo 8* lunar mission, transformed our grasp of its fragile equilibrium. In 2009, we need similarly defining images to galvanize interventions to mitigate climate change. Aiming to provide just that is the exhibition *Earth: Art of a Changing World*, organized by the Royal Academy of Arts in London in collaboration with Cape Farewell, a charity that encourages artists to engage with the science of climate change. It examines how the global-warming debate has influenced the practice of more than 30 international contemporary artists, some of whom have participated in Cape Farewell expeditions to the Arctic.

On entering the exhibition — aptly on show in the building formerly occupied by the now defunct Museum of Mankind — visitors are confronted by the UK sculptor Antony Gormley's *Amazonian Field* (1992), which comprises 15,000 fired clay figures with questioning expressions. "I wanted to make a work about our collective future and our responsibility for it," Gormley has said of his sculpture, hand-modelled by people living in the Amazon basin. In contrast to this cooperative work, the Palestinian sculptor Mona Hatoum's *Hot Spot* (2006) alludes to human conflicts at contested borders. A skeletal globe is tilted at the same angle as Earth, its stainless steel latitudes and longitudes supporting continents outlined in glowing neon tubes. *Hot Spot* hints at the increasing global unrest that could be caused by water shortages resulting from climate change.

Also on show is the work of the UK artists Heather Ackroyd and Dan Harvey, participants in several Cape Farewell expeditions. As part of their ongoing project, Beuys' Acorns (started in 2007), they have planted young oak saplings — grown from acorns collected from trees planted by the pioneering German ecological artist Joseph Beuys in 1982 — on the portico outside the exhibition. Growing more trees, rather than felling the world's forests, is their symbolic and optimistic act in the face of justifiable pessimism about climate change.

"We have come to this ship in a frozen fjord to think about the ways we might communicate our concerns about climate change to a wider public," said UK writer Ian McEwan of the purpose of his and other artists' 2005 Cape Farewell expedition. His latest novel, to be published next year, deals with climate change. For now, his text *The Hot Breath of Our Civilisation* (2005), is exhibited on the gallery wall as a scrolling display. It provides an epilogue to the exhibition: "Are we at the beginning of an unprecedented era of international cooperation, or are we living in an Edwardian summer of reckless denial? Is this the beginning, or the beginning of the end?" ■

**Colin Martin** is a writer based in London.
e-mail: cmpubrel@aol.com

NEWS & VIEWS

EXTRASOLAR PLANETS

# Water world larger than Earth

Geoffrey Marcy

**The hunt for Earth-like worlds has taken a major step forward with the discovery of a planet only 2.7 times larger than Earth. Its mass and size are just as theorists would expect for a water-rich super-Earth.**

Momentous breakthroughs in science often come unexpectedly and serendipitously, requiring decades of patience. Only rarely does a long-sought scientific frontier loom so prominently just beyond the horizon that the next generation of instruments seems sure to reach it. A tantalizing case for such a breakthrough is presented by Charbonneau *et al.*[1] on page 891 of this issue. They provide the most watertight evidence so far for a planet that is something like our own Earth, outside our Solar System.

Charbonneau and his co-workers developed a simple and forward-looking planet-hunting technique. They installed a suite of eight amateur-sized telescopes (with 40-cm-diameter mirrors), each with a sensitive charge-coupled-device light detector that measures the near-infrared brightness (wavelengths of about 700–900 nanometres) of a star. Any star whose brightness dims for about an hour, and repeats that dimming like clockwork over the course of days and weeks, is probably doing so because an orbiting planet is crossing briefly in front of it, blocking a fraction of the star's light. The amount of dimming directly indicates the size of the planet relative to that of the star. From a large sample of nearby stars[2], Charbonneau *et al.* have smartly chosen the 2,000 of smallest radii, so that near-Earth-sized planets would block at least 1% of a star's light, rendering such worlds detectable.

Charbonneau's team[1] has found that the small, faint star GJ 1214 undergoes repeated dimming of 1.3% for 52 minutes every 1.6 days. The only plausible interpretation is that a planet orbits the star with an orbital period of 1.6 days and that it has a radius that is 12% that of the star. Good estimates of the star's radius (21% that of the Sun) put the planet's radius at only 2.7 Earth radii. Such a small planet orbiting a star other than the Sun is an extraordinary find. With the tools currently available, only one other extrasolar planet has been reported that is thought to be close in size to Earth, namely CoRoT-7b, at 1.7 Earth radii. The new planet, which is only about 13 parsecs away, is named GJ 1214b. Importantly, it pulls gravitationally on its host star, causing the star to move with a speed of 12 m $s^{-1}$, which the team has detected through measurements of wavelength shifts in the star's light (the Doppler effect). The planet's inferred mass is a mere 6.6 Earth masses, which, when combined with its radius, leads to a density of 1.9 g $cm^{-3}$. By contrast, Earth's average density is much higher, at 5.5 g $cm^{-3}$. Because water has a low density of about 1 g $cm^{-3}$, the chemical composition of the new planet is probably some admixture of rock and water, with perhaps a small atmosphere of hydrogen and helium.

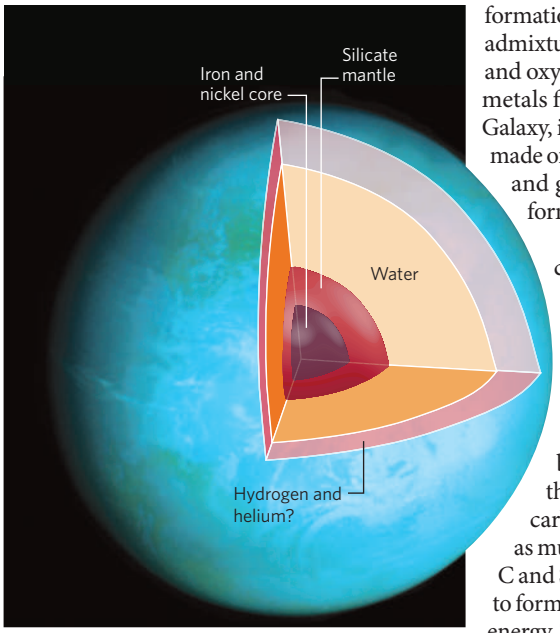



**Figure 1 | A water-rich super-Earth?** The newly discovered[1] extrasolar planet, GJ 1214b, probably contains a huge amount of water, surrounding an inner core of iron and nickel, and an outer mantle of silicate rock, and may have a small atmosphere of hydrogen and helium. Only 2.7 times larger than Earth, and just 13 parsecs away, this super-Earth brings astronomers closer to discovering Earth-like planets.

Could this planet have a solid surface suitable for hosting organic-rich ponds and lakes? Some astronomical background offers a good guess. The protoplanetary disks of dust and gas swirling around young stars are the sites of planet formation. The disks are made up of the same admixtures of H and He gas, carbon, nitrogen and oxygen compounds, and iron and nickel metals found in nearly all of the stars in our Galaxy, including the Sun. Solid dust particles made of Fe, Ni, silicates and ices stick together and grow into ever larger planetesimals, forming the basic cores of all planets.

The relative amounts of these solid constituents vary only modestly among different protoplanetary disks, for two reasons. First, the abundances of the atomic elements are nearly the same, within factors of two, from star to star, with C, N, O, silicon, magnesium and Fe being the building blocks of the solid material. Second, the highly negative Gibbs free energy of carbon monoxide and silicates locks up as much oxygen as the limiting reagents — C and Si — permit, leaving plenty of oxygen to form water ice, despite its higher Gibbs free energy. Thus, silicates and water ice dominate the mass budget of the solid material in the cold regions of protoplanetary disks, along with the Fe and Ni dust grains.

That solid material forms the building blocks of large planets such as Saturn and Neptune, and perhaps smaller planets as well, such as the new one[1]. But the density of 1.9 g $cm^{-3}$ for this new planet imposes a constraint on the relative amounts of each constituent. To keep the planet's density that low requires that it contains large amounts of water. If the planet were pure Fe and silicates, its density would be similar to Earth's. It must contain a huge amount of water, roughly 50% by mass.

The wild card is the amount of H and He gas in the atmosphere. Spooning additional H and He (of low density) onto a planet makes its density lower, which can be compensated for by increasing the amount of Fe in the core to bring the overall density to that measured, 1.9 g $cm^{-3}$. But the planet-building environment is unlikely to spawn planets composed of mostly Fe and H/He, but very little water. Any planet that contains Fe, rock and H/He would have also retained correspondingly large amounts of water. Thus, it is likely that this

new world has nearly 50% of its mass in water surrounding an Fe/Ni core and a silicate mantle (Fig. 1). It probably has an extraordinarily deep ocean, which would be liquid given its equilibrium surface temperature of some 190 °C due to heating from the host star. A sauna-like steam atmosphere is possible, with slow photolytic and hydrodynamic loss of that atmosphere caused by ultraviolet-light irradiation. A thin H/He outer atmosphere is also possible.

And so comes the profound anthropocentric question. If this planet is 50% water, is it really kin of our Earth? Or did it form in a manner similar to that of Saturn or Neptune, with a rocky core that acquired large amounts of ices and gas gravitationally? By contrast, Earth has only 0.06% water, and very little H and He gas, having formed in a dry environment. This new planet is close to Earth in size, but perhaps not next of kin.

Nonetheless, Charbonneau's team[1] has highlighted a promising future for the discovery of Earth-like worlds. Their efforts are just beginning, with smaller and rockier planets yet to be found. Meanwhile, precise Doppler measurements may reveal the gravitational wobbles of stars caused by Earth-like planets in tight orbits. Most promising is NASA's Kepler mission. Launched in March 2009, Kepler is monitoring 100,000 stars and is able to detect dimming as small as one part in ten thousand of their normal brightness, rendering truly Earth-sized planets easily detectable. And someday, great space-borne interferometers (such as NASA's Space Interferometry Mission) and enormous cameras will be launched, able to detect, image and spectroscopically analyse the landscapes, oceans and atmospheres of nearby rocky planets. These techniques will surely answer the question Aristotle, Epicurus and Democritus posed 2,400 years ago regarding Earth's unique status in the Universe. ■

Geoffrey Marcy is in the Department of Astronomy, University of California at Berkeley, Berkeley, California 94720, USA.
e-mail: gmarcy@berkeley.edu

1. Charbonneau, D. *Nature* **462,** 891–894 (2009).
2. Lépine, S. *Astron. J.* **130,** 1680–1692 (2005).

DNA REPLICATION

# Prime-time looping

Nicholas E. Dixon

**When the replication machinery copies DNA, it must unwind the double helix in one direction while synthesis of one of the strands proceeds in the other. Making transient DNA loops may solve this directional dilemma.**

If you are a cell about to divide, you will first need to use a multi-protein machine called a replisome to simultaneously make copies of both strands of your chromosomal DNA so that one strand can be passed to each daughter cell. Replisomes have long been thought to couple synthesis of both DNA strands by forming a 'trombone loop' of DNA that expands and relaxes as synthesis takes place discontinuously on one of the strands. Two papers, one by Pandey *et al.*[1] on page 940 of this issue and another by Manosas *et al.*[2] published in *Nature Chemical Biology*, show that a second type of loop, called the 'priming loop', is transiently produced in the replisome.

The replisome faces special challenges as it makes new DNA at rates that can approach 1,000 nucleotides per second. Unlike the machines that make proteins and RNA, which work relatively sluggishly and in a linear fashion, the replisome must simultaneously copy two strands of DNA that are aligned in opposite directions (5′ to 3′ and 3′ to 5′). Replisome chemistry obeys two rules. The first is that a DNA polymerase (the component of the replisome that synthesizes new DNA from a template strand) can extend the newly formed DNA chain only in the 5′ to 3′ direction. This means that it can continuously copy only one of the two DNA strands, called the leading strand. The lagging strand must be made in shorter pieces that are joined together later. These pieces, or Okazaki fragments, are a few thousand bases in length and each is made every few seconds.

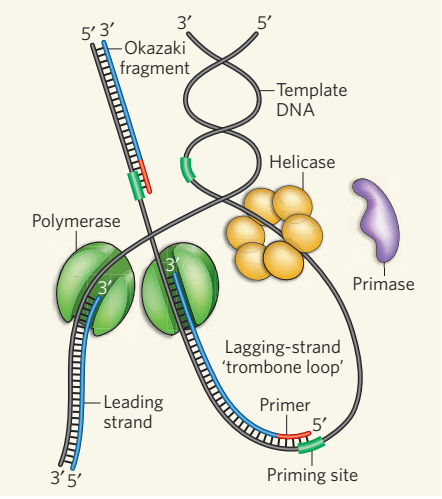



**Figure 1 | DNA replication by a minimal replisome.** During DNA replication by the replisome components, the DNA strands are separated by the helicase enzyme and replicated by the leading- and lagging-strand DNA polymerases. As DNA can be copied only in the 5′ to 3′ direction, the polymerase continuously copies the leading strand, but the lagging strand is made in shorter pieces, or Okazaki fragments, that are joined together later. DNA synthesis begins by extending a nucleic-acid primer that is synthesized at priming sites by the primase enzyme.

The second rule is that a DNA polymerase cannot start a DNA chain — it can only extend a pre-existing DNA or RNA chain, called a primer. So all cells have a specialized enzyme, the primase, that makes the first RNA primer for each DNA chain. A new primer must therefore be made every few seconds to be used for Okazaki-fragment synthesis on the lagging-strand template. This single-stranded template DNA is produced by the helicase, a component of the replisome that, in bacteria, moves in a 5′ to 3′ direction to separate the two strands of the double helix (Fig. 1). And herein lies the problem — the primase needs to be associated with the helicase to function, but the primers on the lagging strand are made in the direction opposite to the movement of the helicase. Moreover, primer synthesis is relatively sluggish, taking about a second or so.

There are three possible solutions to the replisome's problem. One is for the whole replisome to pause while the primer on the lagging strand is made, then to resume its work; such pauses have been reported by the van Oijen group[3] during primer synthesis by the bacterial virus (bacteriophage) T7 replisome (Fig. 2a). The second solution is for the primase, once clamped onto the lagging-strand template by the helicase, to be promptly released to make its primer at leisure, as happens with the *Escherichia coli* replisome[4] (Fig. 2b).

The third solution is for the replisome to continue leading-strand synthesis while the helicase–primase complex takes its time to make the primer. The helicase continues to unwind DNA in the forward direction while the physically linked primase makes a primer in the opposite direction. This arrangement produces a transient single-stranded DNA loop in the lagging-strand template, termed the priming loop, which is subsequently released to become part of the trombone loop when the primer is passed to the lagging-strand polymerase (Fig. 2c).

The new reports[1,2] use elegant single-molecule experiments to provide the first direct experimental evidence for priming-loop formation by the bacteriophage T7 and T4 replisomes. Pandey *et al.*[1] worked with the whole T7 replisome, which has an unusual structure in that its primase and helicase are part of the same protein, so primase release is impossible. The authors used short DNA templates that were already primed on the leading strand, with priming sites (DNA sequences required for primer synthesis) on the lagging strand. Although lagging-strand primer synthesis occurred about 50% of the time, synthesis of the leading strand showed no sign of pausing while a primer was made. Next, the authors[1] employed a technique called fluorescence

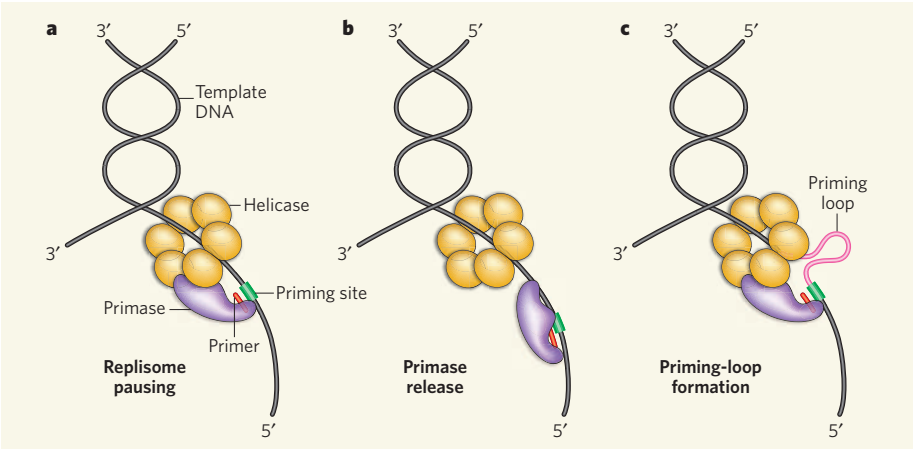



**Figure 2 | Three priming mechanisms.** Interaction of the primase with the helicase is necessary for primer synthesis at a lagging-strand priming site. The primase makes primers in the opposite direction to helicase movement, leaving three ways by which the replisome might resolve this directional problem. **a**, The whole replisome can pause for primer synthesis; **b**, it can promptly release the primase; or **c**, as described for the first time by Pandey *et al.*[1] and Manosas *et al.*[2], the replisome can continue to move forward while the primase–helicase interaction persists. This produces a priming loop that eventually collapses into the lagging-strand trombone loop, probably on transfer of the primer to the lagging-strand polymerase.

resonance energy transfer (FRET), which uses the interaction between fluorescent dyes as a readout of the proximity of molecules to each other. The dyes were arranged on the lagging-strand template so that they would come close enough together for FRET to occur if a priming loop were formed. FRET was observed only under conditions where, and about as often as, primers were made. The FRET data[1] can be explained only by the formation of a priming loop on the lagging-strand template while leading-strand synthesis continues (Fig. 2c).

In another single-molecule study, Manosas *et al.*[2] studied the T4 replisome, in which the primase and helicase are separate proteins that interact during primer synthesis. They used an ingenious experimental design consisting of a double-stranded DNA hairpin structure that contains priming sites when in a single-stranded form. The DNA is attached to a magnetic bead that is stretched at a constant low force by a magnetic field. Videomicroscopy of the bead movement allows measurement of the length of the DNA. As the helicase converts the hairpin to single-stranded DNA, the DNA lengthens and then subsequently contracts as the hairpin reanneals behind it. The changes in DNA length allow measurement of the rate of helicase action in real time. Using this system, the authors[2] showed that helicase–primase interaction and subsequent primer synthesis did not result in helicase pausing. Most of the time, reannealing of the hairpin was blocked by the persistence of a primase-bound primer, indicating that the primase had been released promptly by the helicase at the priming site (Fig. 2b). Less frequently, the rate of DNA lengthening decreased for about half a second, and then there was an immediate jump in length. This observation can be explained only by the formation and subsequent release of a priming loop (Fig. 2c). When the helicase and primase were artificially fused together as in the T7 replisome, priming-loop formation was markedly increased, and blocks to reannealing (by released primase-bound primer) were not observed.

An unusual aspect of Pandey and colleagues' work[1] is the high efficiency of priming achieved by the T7 primase on their short templates. Priming sites are trinucleotides that occur frequently in single-stranded DNA templates. They are generally used inefficiently by the primase for primer synthesis, and it is thought that only a fraction of primers are functionally extended by the lagging-strand polymerase. These factors account for the relatively long (1–2 kilobases) Okazaki fragments. When studying lagging-strand priming during leading-strand synthesis by the T7 replisome on long templates, the van Oijen group[3] clearly observed pauses coincident with primer synthesis. These occurred at relatively low frequency, consistent with the size of Okazaki fragments — but the authors' single-molecule experimental set-up could not detect priming loops. Reconciliation of these observations[3] with those of Pandey *et al.*[1] is not straightforward, and may indicate that replisome pausing occurs during or soon after functional primer synthesis, while mechanisms involving primase dissociation and priming-loop formation ensure that the replisome is not unnecessarily slowed during more frequent, non-productive priming events. ■

Nicholas E. Dixon is at the School of Chemistry, University of Wollongong, Wollongong, New South Wales 2522, Australia.
e-mail: nickd@uow.edu.au

1. Pandey, M. *et al. Nature* **462,** 940–943 (2009).
2. Manosas, M. *et al. Nature Chem. Biol.* **5,** 904–912 (2009).
3. Lee, J.-B. *et al. Nature* **439,** 621–624 (2006).
4. Yuzhakov, A., Kelman, Z. & O'Donnell, M. *Cell* **96,** 153–163 (1999).




## 50 YEARS AGO

*A Survey of Soils in the Kongwa and Nachingwea Districts of Tanganyika*. By B. Anderson — Everyone knows the dismal sequel to the ambitious scheme for the mechanized production of ground-nuts in East Africa, which was characterized by the failure to employ pedological methods on pedological problems. After the horse had departed, some effort was made to shut the stable door and a very competent soil surveyor was set to work to make a proper study of the soils of the Kongwa and Nachingwea districts. The publication under review presents his results and shows what can be achieved by one trained pedologist working 'on the cheap' with limited facilities, but with specialized technical assistance from various institutions. The moral for would-be planners of land-use is obvious.
From *Nature* 19 December 1959.

## 100 YEARS AGO

It may be of interest to record a fact which has come under my notice while engaged in the development of a uranious mine in Turkestan. The ore is oxidised and calcareous, and contains uranium, vanadium, and copper, radium being present in accordance with Prof. Rutherford's formula, which gives the quantity of it in relation to the uranium. The uranium is on the average 3.8 per cent., but in some places reaches the ratio of 30 per cent. and more ... As I know from the literature of the subject that vanadium and uranium are toxic substances, I instruct the workmen to wash their hands well before going to their dinner and after their work. "We do this," they say, "but at the same time we know that in actual practice a cut on a hand, which lasts for a long time in a coal mine, here, when powdered by the ore, gets well very quickly."
From *Nature* 16 December 1909.

50 & 100 YEARS AGO

GLOBAL CHANGE

# Interglacial and future sea level

Peter U. Clark and Peter Huybers

**A merger of data and modelling using a probabilistic approach indicates that sea level was much higher during the last interglacial than it is now, providing telling clues about future ice-sheet responses to warming.**

Predicting sea-level rise in a warming world is one of science's great challenges. According to sea-rise projections for the twenty-first century, the 145 million people living within a metre of the present sea level risk losing their land and their homes. Many more would be affected by the resulting socio-economic disruption[1]. Our poor understanding of ice-sheet dynamics means that projecting sea-level rise beyond the twenty-first century is much less certain[2]. On page 863 of this issue, however, Kopp *et al.*[3] derive a new assessment of sea level during the last interglacial, around 125,000 years ago, that provides insight into this question. If their results are correct, the sea-level rise over the coming century will be followed by many more metres of rise over the ensuing centuries.

Increases in global sea level stem from both expansion of warming water (thermosteric change) and addition of new water from melting ice on land (eustatic change). Predictions of future thermosteric changes are relatively well constrained compared with those of the eustatic change associated with melting of the Greenland and Antarctic ice sheets[4]. There is thus a need to better determine both how much and how rapidly eustatic sea level will rise in response to a given forcing effect such as anthropogenic global warming.

Evidence that sea level during the last interglacial was 4–6 metres higher than at present has long been proposed as a possible analogue for the equilibrium sea-level response to future anthropogenic warming[5,6]. But the sea-level records may include a local response to geophysical adjustments from the preceding glaciation, and thus may not accurately record the global sea level[7]. Furthermore, the implications of 4 or 6 m of rise are quite different: if sea level increases by only 4 m, much of it can be reconciled as being due to thermosteric rise and partial loss of the Greenland ice sheet; anything more requires a contribution from Antarctica.

Kopp *et al.*[3] reach the startling conclusion that, during the last interglacial, global sea level was at least 6.6 m above present, and may have reached 9.4 m, much higher than previous estimates. The implication is that both the Greenland and Antarctic ice sheets were much smaller 125,000 years ago.

To derive this result, Kopp *et al.* compiled a database of proxy measurements of sea level that includes isotopic and coral records, as well as other records that are less well dated. Although this database is more comprehensive than those used in previous studies, constraining estimates for past global sea level from noisy and sparse data whose timing is uncertain is a formidable statistical problem. It is particularly difficult because one must also account for regionally varying geophysical effects, including local tectonic uplift or subsidence, and sea-level changes induced by gravitational, deformational and rotational effects associated with the redistribution of ice, ocean and mass of the solid Earth[8]. Using a physical model that includes these effects, Kopp *et al.* derived an estimate of the covariance between local and global sea level. They then merged the local–global covariance estimate with proxy estimates of sea level within a Bayesian framework to make temporally complete estimates of global sea level and assess their probability.

The redistribution of mass associated with individual ice-sheet melting causes distinct spatial patterns in sea level[9]. In conjunction with the proxy measurements, Kopp *et al.*[3] also used the modelled patterns to estimate that Greenland and Antarctica each contributed at least 2.5 m of sea-level rise. This estimate is consistent with independent constraints: the maximum Greenland contribution was probably 3.4 m (ref. 10), and the thermosteric plus mountain-glacier and ice-cap contribution was probably no more than 1 m. So, if sea level was at least 6.6 m higher, a minimum of 2.2 m must have come from Antarctica. The Antarctic contribution would probably have come from the inherently unstable West Antarctic Ice Sheet, which locks up the equivalent of at least 3.3 m of sea level[11], so that Kopp and colleagues' result implies that most, if not all, of this ice sheet melted about 125,000 years ago.

Perhaps of greatest socio-economic concern is the possible maximum rate of sea-level rise in a warmer world. According to Kopp *et al.*[3], sea-level rise during the last interglacial was in the range of 6–9 millimetres per year. By comparison, instrumental records indicate that the rate of global sea-level rise over the twentieth century was about 2 mm $yr^{-1}$. That may have accelerated between 1993 to 2003 to around 3 mm $yr^{-1}$, at least in part due to an acceleration in mass loss from the Greenland and Antarctic ice sheets[12].

Why was sea level so much higher 125,000 years ago? One possibility is that ice sheets have multiple potential steady states for a given climate[13]. However, the global temperature was apparently 1.5–2 °C warmer than the pre-anthropogenic global average of the past 10,000 years (Fig. 1), despite there being essentially no difference in atmospheric greenhouse-gas concentrations. Climate models have simulated a strong Northern Hemisphere summer warming in response to Earth's more eccentric orbit during the last interglacial, but almost no change in the Southern Hemisphere[14]. Southern warming may then have occurred through an oceanic teleconnection with the north[15], or through changes in the duration of the Southern Hemisphere summer[16], with accompanying feedbacks amplifying this warming.

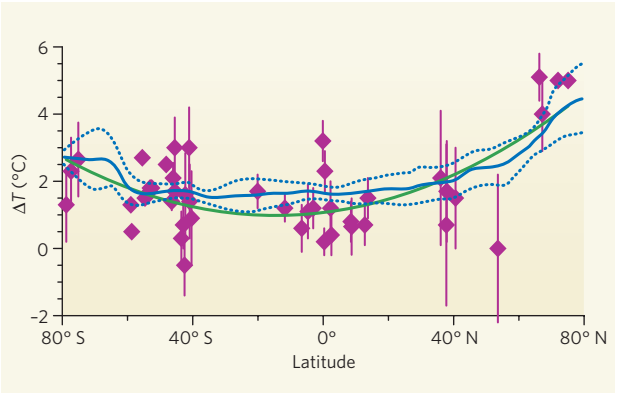



**Figure 1 | Similarity of latitudinal warming (Δ*T*) during the last interglacial and a projection for the late twenty-first century.** The green line summarizes proxy-data estimates of sea surface and air temperature during the last interglacial relative to the present interglacial before industrialization. Diamonds are largely sea surface temperatures, but include temperatures derived from polar ice cores and two high-latitude Northern Hemisphere pollen records. The temperatures reflect the interval between 120,000 and 130,000 years ago (mean and 1 standard deviation). The green line is a polynomial fit to these data. Surface air temperature estimates from less-well-dated pollen sites in Europe (not shown) similarly show warmer temperatures across most of Europe during the last interglacial[17]. The blue solid line is the zonal mean of the projected surface temperature changes (with 1 standard deviation shown by dotted blue lines) for the late twenty-first century relative to 1980–99; it is based on the SRES B1 greenhouse-gas-emission scenario obtained using the GFDL climate model. (Palaeoclimate data are available at www.ncdc.noaa.gov/paleo/pubs/clark2009.)

In any event, the latitudinal distribution of warming seems to be remarkably similar to the global temperature response to carbon dioxide under a commonly used scenario for greenhouse-gas emissions (compare the green and blue lines in Fig. 1). This suggests that the climate of the last interglacial might, by coincidence, provide a reasonable analogue for establishing ice-sheet sensitivity to global warming. Assuming that Kopp and colleagues' estimates are accurate, and that higher sea level resulted from higher temperatures, the disconcerting message is that the

equilibrium response of sea level to 1.5–2 °C of global warming could be an increase of 7–9 metres. ■

Peter U. Clark is in the Department of Geosciences, Oregon State University, Corvallis, Oregon 97331, USA. Peter Huybers is in the Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mails: clarkp@onid.orst.edu; phuybers@fas.harvard.edu

1. Anthoff, D., Nicholls, R. J., Tol, R. S. J. & Vafeidis, A. *Tyndall Centre Working Pap. 96* (2006).
2. Alley, R. B., Clark, P. U., Huybrechts, P. & Joughin, I. *Science* **310,** 456–460 (2005).
3. Kopp, R. E., Simons, F. J., Mitrovica, J. X., Maloof, A. C. & Oppenheimer, M. *Nature* **462,** 863–867 (2009).
4. Meehl, G. A. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. D. *et al.*) 747–845 (Cambridge Univ. Press, 2007).
5. Mercer, J. H. *Nature* **271,** 321–325 (1978).
6. Jansen, E. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. D. *et al.*) 433–497 (Cambridge Univ. Press, 2007).
7. Lambeck, K. & Nakada, M. *Nature* **357,** 125–128 (1992).
8. Mitrovica, J. X. & Milne, G. A. *Geophys. J. Int.* **154,** 253–267 (2003).
9. Mitrovica, J. X., Tamisiea, M. E., Davis, J. L. & Milne, G. A. *Nature* **409,** 1026–1029 (2001).
10. Otto-Bliesner, B. L. *et al. Science* **311,** 1751–1753 (2006).
11. Bamber, J. L., Riva, R. E. M., Vermeersen, B. L. A. & LeBrocq, A. M. *Science* **324,** 901–903 (2009).
12. Velicogna, I. *Geophys. Res. Lett.* doi:10.1029/2009GL040222 (2009).
13. Pollard, D. & DeConto, R. M. *Global Planet. Change* **45,** 9–21 (2005).
14. Crowley, T. J. & Kim, K.-Y. *Science* **265,** 1566–1568 (1994).
15. Duplessy, J. C., Roche, D. M. & Kageyama, M. *Science* **316,** 89–91 (2007).
16. Huybers, P. & Denton, G. *Nature Geosci.* **1,** 787–792 (2008).
17. Kaspar, F., Kühl, N., Cubasch, U. & Litt, T. *Geophys. Res. Lett.* doi:10.1029/2005GL022456 (2005).

DNA REPAIR

# A heavyweight joins the fray

Simon J. Boulton

**Tagging of DNA-damage-associated proteins by ubiquitin is key to coordinating the DNA-damage response. The ubiquitin-related protein SUMO is revealed as a crucial regulator of ubiquitylation in DNA repair.**

Ubiquitylation — the attachment of ubiquitin groups to cellular proteins — was initially characterized by its role in promoting protein destruction. However, we now know that the consequences of ubiquitylation are diverse, and that it affects many cellular systems. The ubiquitin modification comes in many flavours (addition of a single ubiquitin molecule, for example, or of polyubiquitin chains that differ in the position of the linkage between ubiquitin molecules), and the various types of ubiquitylation can alter the fate of target proteins in different ways. In addition, the cell has ubiquitin-related modifiers, such as the SUMO proteins, that also alter protein fate or function after conjugation[1]. One process that has been inextricably linked to ubiquitylation is the cellular response to DNA damage. Although studies[2,3] had suggested a link between the DNA-damage response and the SUMO pathway, proof that SUMOylation is important for DNA repair had remained elusive. In this issue, two groups, Morris *et al.*[4] (page 886) and Galanty *et al.*[5] (page 935), now provide good evidence that SUMO functions together with ubiquitin to coordinate DNA repair.

DNA double-strand breaks (DSBs) result in the recruitment and activation of the protein kinases ATM, ATR and DNA-PK, which phosphorylate target proteins, such as the variant histone H2AX. The phosphorylated proteins then promote the recruitment of other DNA-repair proteins to DSBs[6], including MDC1 (mediator of the DNA-damage checkpoint), 53BP1 and the E3 ubiquitin ligases RNF8, RNF168 and BRCA1 (ref. 6), which catalyse ubiquitylation events[7] at DSBs. (Conjugation of ubiquitin or related modifiers to target proteins requires an E1 activating enzyme, an E2 conjugating enzyme and an E3 ligase.)

To investigate the involvement of the SUMO pathway in the DNA-damage response, Morris *et al.*[4] and Galanty *et al.*[5] analysed the subcellular localization of SUMO-pathway components in mammalian cells. Both groups[4,5] report that the E1 SUMO-activating enzyme SAE1, the E2 SUMO-conjugating enzyme UBC9, and the three forms of vertebrate SUMO protein, SUMO1 and the closely related SUMO2 and SUMO3 (SUMO2/3), are recruited to DSBs.

The authors[4,5] used RNA interference and fluorescence microscopy to show that the SUMO E3 ligases PIAS1 and PIAS4 are responsible for SUMOylation events at DSBs. Depletion of PIAS1 impaired accumulation of SUMO2 and SUMO3 (but not SUMO1) at DSBs, whereas depletion of PIAS4 impaired recruitment of SUMO1 and SUMO2/3. Furthermore, recruitment of 53BP1 to DSBs depended on PIAS4, whereas recruitment of BRCA1 depended on both PIAS1 and PIAS4. Is SUMOylation necessary for DSB repair? The answer is, emphatically, yes — cells lacking PIAS1 or PIAS4 showed defects in DSB repair and were also highly sensitive to DSBs caused by ionizing radiation.

What are the targets of the SUMO pathway during the DNA-damage response? Prompted by a study showing interaction between UBC9 and BRCA1 in the nematode worm *Caenorhabditis elegans*[2], both groups[4,5] independently showed that BRCA1 is SUMOylated during the DNA-damage response in a PIAS1- and PIAS4-dependent manner (Fig. 1). Depletion of PIAS1 and PIAS4 impaired recruitment of BRCA1 to DSBs[4,5], significantly impaired ubiquitylation at DSBs, and reduced ubiquitylation of the histones H2A and H2AX; the latter process has been shown to require the ligase activities of RNF8, RNF168 and BRCA1 (ref. 7). Galanty *et al.*[5] also showed that 53BP1 is SUMOylated and that this affects its retention at DSBs.

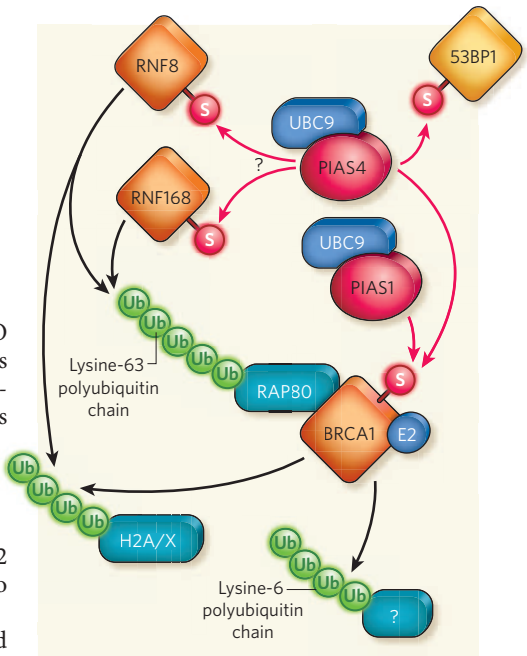



**Figure 1 | Ubiquitylation and SUMOylation at DSBs.** Double-strand DNA breaks (DSBs) result in the recruitment of DNA-repair proteins, including 53BP1 and the E3 ubiquitin ligases RNF8, RNF168 and BRCA1. Morris *et al.*[4] and Galanty *et al.*[5] observe that the SUMO-pathway components UBC9–PIAS4 and UBC9–PIAS1 also accumulate at DSBs, where they catalyse the SUMOylation of 53BP1 and BRCA1 (and possibly RNF8 and RNF168). SUMOylation stimulates BRCA1 E3 ubiquitin-ligase activity, leading to ubiquitylation of target proteins at DSBs, including the histone H2A and its variant H2AX. H2A and H2AX are also substrates for ubiquitylation by RNF8 and RNF168, as is RAP80, a ubiquitin-binding protein that also interacts with BRCA1. RNF8 and RNF168 catalyse the formation of lysine-63-linked ubiquitin chains, whereas BRCA1 and its E2 conjugating enzyme catalyse the formation of lysine-6-linked ubiquitin chains. S, SUMO; Ub, ubiquitin. Red arrows indicate SUMOylation; black arrows indicate ubiquitylation.

RNF8 and RNF168 catalyse the formation

of lysine-63-linked ubiquitin chains, whereas BRCA1 catalyses formation of lysine-6-linked ubiquitin chains[8] (Fig. 1). Morris *et al.*[4] exploited this difference in ubiquitin-chain linkage to pinpoint the effects of PIAS proteins on BRCA1 activity. They showed that overexpression of BRCA1 increased ubiquitylation events in cells; these events were reduced following PIAS1/4 depletion. Co-localization of lysine-6-linked ubiquitin chains with DSBs was also impaired in BRCA1-, PIAS1- or PIAS4-depleted cells. Furthermore, mutation of the two consensus SUMO-conjugation sites in BRCA1 reduced SUMO1 association and BRCA1-dependent ubiquitylation. Thus, the authors propose that BRCA1 is a SUMO-regulated ubiquitin ligase (Fig. 1).

These findings raise several immediate questions. Are the activities of RNF8, RNF168 and/or other E3 ubiquitin ligases also regulated by SUMOylation? Certainly, such a scenario is possible for RNF8. The current studies[4,5] found that, although recruitment of RNF8 to DSBs was unaffected by PIAS1/4 depletion, RNF8 could not ubiquitylate DSBs, suggesting that it may be inactive in the absence of PIAS1/4. How does SUMOylation stimulate the E3 ubiquitin-ligase activity of BRCA1? Previous studies[9] have shown that DNA damage promotes association between BRCA1 and its E2 conjugating enzyme to form an active E3 ubiquitin ligase. It is therefore tempting to speculate that SUMOylation induces a conformational change in BRCA1 that enhances its binding to an E2 conjugating enzyme. It is clear from the current studies that SUMOylation functions at multiple levels during the DNA-damage response and this will provide fertile ground for future research. The discovery that the SUMO pathway is important for ubiquitylation at DSBs raises the possibility that SUMOylation may activate other ubiquitylation events in the cell. ■

Simon J. Boulton is at the DNA Damage Response Laboratory, London Research Institute, Cancer Research UK, Clare Hall, South Mimms EN6 3LD, UK.
e-mail: simon.boulton@cancer.org.uk

1. Welchman, R. L., Gordon, C. & Mayer, R. J. *Nature Rev. Mol. Cell Biol.* **6,** 599–609 (2005).
2. Boulton, S. J. *et al. Curr. Biol.* **14,** 33–39 (2004).
3. Golebiowski, F. *et al. Sci. Signal.* **2,** ra24 (2009).
4. Morris, J. R. *et al. Nature* **462,** 886–890 (2009).
5. Galanty, Y. *et al. Nature* **462,** 935–939 (2009).
6. Harper, J. W. & Elledge, S. J. T. *Mol. Cell* **28,** 739–745 (2007).
7. Panier, S. & Durocher, D. *DNA Repair* **8,** 436–443 (2009).
8. Morris, J. R. & Solomon, E. *Hum. Mol. Genet.* **13,** 807–817 (2004).
9. Polanowska, J., Martin, J. S., Garcia-Muse, T., Petalcorin, M. I. & Boulton, S. J. *EMBO J.* **25,** 2178–2188 (2006).

NANOTECHNOLOGY

# Soggy origami

Vincent H. Crespi

**Flat microstructures can be designed to spontaneously fold into three-dimensional shapes. Computer simulations of water droplets on sheets of carbon atoms now extend this concept to the nanometre scale.**

Explanations of nanometre-scale phenomena often require strange bedfellows of scientific concepts and terminology. The work reported by Patra *et al.*[1] in *Nano Letters* nicely illustrates this trend by marrying chemistry, fluid mechanics, mechanical engineering and physics. The authors have used molecular dynamics simulations to show that the catalytic action of nanodroplets of fluids can cause a simple object — an atomically thin layer of carbon atoms, known as a graphene sheet — to fold spontaneously into complex shapes. The ramifications of this scientific polygamy extend beyond the four fields mentioned above: such spontaneous folding evokes the behaviour of proteins, and graphene sheets also hold promise for electronic applications.

Graphene sheets are single layers of graphite (the familiar stuff in pencils), in which the carbon atoms are arranged in a honeycomb pattern. When Patra *et al.* simulated a small droplet of water sitting on such a sheet, they found that, rather than simply sitting still, the droplet actively deforms the graphene membrane. Because atoms at exposed surfaces in materials are less stable than those buried deeper within, the authors' droplet–membrane system collectively deforms to minimize the number of exposed atoms and hence lower the system's energy. The precise deformation depends on the shape of the graphene sheet and the diameter of the droplet, but, in general, the sheet wraps up around the droplet to form folds, scrolls or troughs (Fig. 1a). When a droplet interacts with a sheet shaped like an open, four-petalled flower, the petals fold up around the droplet to form a closed bulb (Fig. 1b).

In principle, a graphene membrane without a droplet could also lower its energy by bending, so that some of the surface carbon atoms become buried in the interior of a more compact shape. But the transition of an exposed, flat membrane into a smaller, folded package is problematic for an isolated sheet: the energy cost of forming intermediate, partially curled shapes is not compensated for by the short-range attraction between the approaching surfaces until the sheet has bent enough for the surfaces to touch.

This is where the nanodroplets come in. Patra and colleagues' study[1] shows that fluid droplets act as catalysts for graphene deformation — they remove the energy barrier that prevents folding reactions, without themselves undergoing any structural changes. Remarkably, after a droplet has done the work of folding the sheet, it can be expelled from the resulting structure as the opposing graphene surfaces press tightly against each other. Apparently, graphene surfaces prefer being wet to being naked, but they prefer the contact of other graphene surfaces even more. As a result, fluid nanodroplets can cleanly convert flat graphene sheets into folded bilayers, then leave gracefully after their work is done. Such bilayered graphene systems are interesting in their own right as they profoundly alter the remarkable electronic properties of graphene[2].

Spontaneous curling or folding is not unique to soggy graphene. It is also seen in a variety of other systems, ranging from micromachines to biomolecules. The spontaneous deformation of flat sheets was a nuisance in early work on micromachines that were etched from silicon or related materials. When thin sheets of material were released from an underlying substrate, they would curl up in an uncontrolled

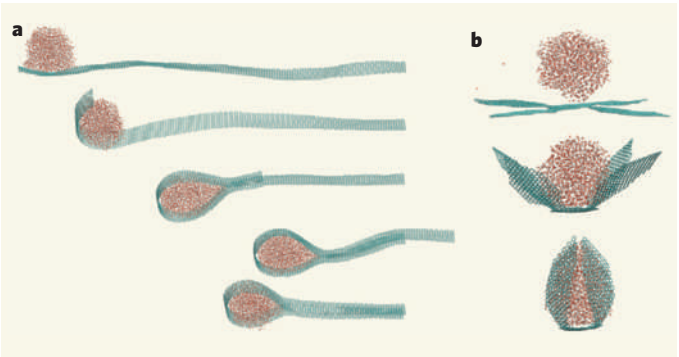



**Figure 1 | Folding sheets.** Patra and colleagues' molecular dynamics simulations[1] reveal that droplets of water cause atomically thin layers of graphite (graphene sheets) to fold up into more compact shapes. **a,** This simulation shows that a nanodroplet of water molecules (red) causes the free end of a graphene ribbon 30 × 2 nanometres in size to wrap around it. The two opposing sheets then slide along each other, folding the ribbon in half. **b,** Here, a droplet of water causes the 'petals' of a flower-shaped graphene ribbon about 13 nanometres across to fold up around it. (Images taken from ref. 1.)

fashion to release previously hidden internal stresses in the material, thus destroying the carefully planned geometry of the desired machine.

The unwanted deformations of micromachines were subsequently brought under control, and even turned to good use, by deliberately engineering stresses into materials to generate a preferred direction for bending. This can be achieved by considering the atomic structures of the materials. Every crystalline solid has its own preferred spacing between constituent atoms — the atoms in germanium, for example, are more widely spaced than those in silicon. If two sheets of different crystalline materials are layered to form a thin bilayer sheet in which the atoms across the interface are mutually aligned, then the layer that prefers a larger inter-atomic spacing is placed under compression, whereas the other layer is placed under tension. These internal strains can be relaxed if the sheet spontaneously curls away from the compressed side. Such systems can be designed to bend or curl into desired shapes, such as scrolls, spirals and even pop-up structures[3,4]. By contrast, there is no way to build such stresses into a single layer of graphene. Patra and colleagues' findings[1] circumvent this problem: the interactions of the graphene sheet with the liquid drop define the way that the sheet will curl.

The interactions of fluids with materials have also been exploited at the micrometre scale to create spontaneously folding structures. In these cases, the size of the structures allows the use of photolithography — a finely honed technique best known for sculpting integrated circuits out of silicon — to precisely define the geometry of initially flat shapes, which subsequently fold when regions of solder within them are melted. The natural tendency of the solder droplets is to minimize their exposed surface area; in doing so, they induce the structures to pull themselves into more compact, three-dimensional objects such as cubes or tetrahedra[5]. Patra and colleagues' simulations[1] potentially extend this ingenious technique to objects a hundred-fold smaller.

The three-dimensional structures of polymeric biomolecules, such as proteins and DNA, are formed by similar, exquisitely precise folding of one-dimensional chains. Humans have learned to exploit this phenomenon, particularly in the practice of DNA origami, wherein specific interactions between complementary DNA strands are programmed to interweave a backbone helix into a desired shape through the incorporation of so-called staple strands[6,7]. The folding of proteins, by contrast, is governed at the crudest level by the tendency of hydrophobic (water-repellent) regions to curl up within the interior of folded protein structures. In this way, hydrophobic protein domains become shielded from their watery environment by hydrophilic (water-attracting) regions of the same protein strand. Could hydrophobic or hydrophilic groups be engineered into graphene to modulate its folding, so extending the one-dimensional lessons of biology to the two-dimensional world of graphene? The jury is still out, but Patra and colleagues' preliminary work[1] certainly opens up investigations of this idea.

The quantitative details of Patra and co-workers' empirical simulations — particularly those concerning subtle sheet–fluid and sheet–sheet interfacial interactions — merit verification by more precise methods. For many practical applications, graphene would lie on a substrate, and so it would also be useful to incorporate sheet–substrate interactions into a second generation of simulations. But the fundamental physics described by Patra and colleagues' models is undoubtedly correct. Experimental validation of their findings is the next obvious step. The availability of a wide assortment of fluids suggests that the physical balance of fluid–sheet and fluid–fluid interactions required to bring about graphene origami should be possible in the real world. ■

Vincent H. Crespi is in the Departments of Physics and Materials Science and Engineering, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: crespi@phys.psu.edu

1. Patra, N., Wang, B. & Král, P. *Nano Lett.* **9,** 3766–3771 (2009).
2. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. *Rev. Mod. Phys.* **81,** 109–162 (2009).
3. Allen, J. J. *Micro Electro Mechanical System Design* (CRC Press, 2005).
4. Cho, A. *Science* **313,** 164–165 (2006).
5. Leong, T., Gu, Z., Koh, T. & Gracias, D. H. *J. Am. Chem. Soc.* **128,** 11336–11337 (2006).
6. Aldaye, F. A., Palmer, A. L. & Sleiman, H. F. *Science* **321,** 1795–1799 (2008).
7. Rothemund, P. W. K. *Nature* **440,** 297–302 (2006).

NEUROSCIENCE

# New tricks and old spines

Noam E. Ziv and Ehud Ahissar

**Imaging of brain structures in living mice reveals that learning new tasks leads to persistent remodelling of synaptic structures, with each new skill associated with a small and unique assembly of new synapses.**

The notion that structural changes in brain circuitry underlie certain forms of learning is widely accepted, yet this belief has been frustratingly difficult to establish experimentally. Two studies, one by Xu *et al.*[1] (page 915) and one by Yang *et al.*[2] (page 920) published in this issue, provide compelling evidence that learning new motor tasks (and acquiring new sensory experiences) is associated with the formation of new sets of persistent synaptic connections in motor (and sensory) regions of the mouse brain. These findings suggest that synapse assemblies, rather than cell assemblies, might be viewed as the elementary entities (engrams) of stored memories.

At a basic level, the brain can be viewed as a vast network of neurons connected to each other by specialized structures known as synapses. Most synaptic connections are formed between axons (the slender and elongated extensions that carry the signals generated by neurons) and dendrites (highly branched extensions that are specialized for receiving signals originating in other neurons). Most axodendritic synapses are found on dendritic spines, which are tiny protrusions that extend from dendritic shafts in large numbers (Fig. 1a, overleaf).

It has long been assumed that structural changes in this complex circuitry provide the basis for long-term memory formation or the learning of new tasks[3]. The development of new imaging tools[4] over the past decade has opened the door to an experimental evaluation of this assumption as it has allowed for longitudinal studies of axonal and dendritic morphology in the brain of living animals (mainly mice). Somewhat surprisingly, it has become evident that overall neuronal morphology is remarkably stable over long periods (months and beyond). What such studies have revealed, however, is that structural changes do occur at the level of individual synapses, manifested by the appearance of new dendritic spines and the disappearance of others over the course of hours and days[3,5,6]. Interestingly, the extent of spine remodelling can be altered by experimental manipulations, such as depriving the animals of sensory input (from the whiskers or eyes)[3,5,6]. Yet a direct relationship between spine remodelling and learning had yet to be demonstrated.

The two studies in this issue[1,2] provide strong evidence for this link. In both studies, mice were trained to perform a new motor task (reaching for a single seed[1] or remaining on an accelerating rotating rod[2]), and two-photon microscopy[4] of the living animals was used to investigate whether successful training was associated with changes in spine remodelling beyond those observed in untrained mice. Both studies revealed that by the end of the first 1–2 days of training (and as soon as one hour after training[1]), twice as many new spines had appeared in the brain of trained mice compared with untrained mice. Continuous training was subsequently followed by increased rates of spine elimination, and so, after 1–2 weeks,

NATURE|Vol 462|17 December 2009

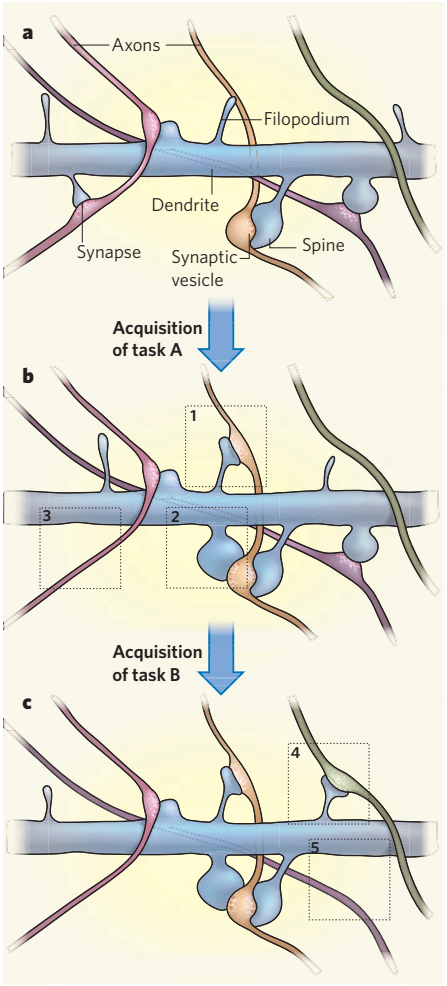



**Figure 1 | Structural remodelling associated with motor task learning[1,2].** **a**, Several axons can form *en passant* synapses (synapses formed along axons rather than at their tips) with dendritic spines extended from a nearby dendrite. Synaptic vesicles contain neurotransmitters secreted at these synapses. **b**, The acquisition of new motor task A is associated with the formation of new spines. This can occur by the selective stabilization of a dendritic filopodium that had contacted a nearby axon beforehand (spine 1) or by the direct emergence of a new spine that extends towards a nearby axon (spine 2). Motor task acquisition is then followed by the elimination of other spines (3). **c**, Acquisition of a second motor task (task B) is followed by the formation of an additional set of new spines (4) and the elimination of others (5), without significantly affecting the spines formed during the acquisition of task A.

total spine numbers did not differ between trained and untrained animals. Remarkably, behavioural performance correlated well with the numbers of new spines formed shortly after training (and with the extent of pre-existing-spine elimination) when these parameters were compared across all trained animals.

Yang and colleagues[2] also examined how exposure to an enriched environment (altering patterns of bead strings hanging from cage tops) affected spine remodelling. In essence, spine remodelling was affected in the same manner. Here, however, remodelling was confined to particular regions of the cerebral cortex concerned with sensory input from the whiskers, unlike spine remodelling associated with motor task learning, which was confined to specific cerebral cortex regions concerned with forelimb movement[1,2].

Although these findings strongly support the notion that some forms of learning are 'encoded' by changes in brain circuitry, several points warrant further discussion. In agreement with other studies[3,5,6], most (96–98%) of the new spines in both trained and untrained animals were short-lived (days), with <1% persisting for many months[2]. Therefore, only a tiny fraction of the new spines formed over a 2-day period would be expected to survive for long durations. Yang and colleagues[2] calculated that at the end of the mouse's life (assumed to be 36 months) these new spines would make up 0.04% of the total number of spines in the particular cortical areas examined. Although this tiny fraction is suggested to be a sufficiently large number of spines to encode a learned behaviour, one wonders how baseline changes in spine number, shown in the same studies[1,2] to be about 3–5% per day (two orders of magnitude greater), can occur without some loss of coherent brain function.

One possible explanation might be that most short-lived spines are only weakly functional[7,8]. It is intriguing to speculate that this large pool of weak synapses could serve as a substrate for selective processes that would promote the stabilization and maturation of a minority of such synapses according to certain functionality criteria. These synapses would then survive and eventually underlie future behaviour. In this respect it is worth mentioning that training was not associated with the additional proliferation of dendritic filopodia[1,2] (often considered to represent spine precursors[9]) but was associated with increased conversion of filopodia to spines (Fig. 1b, spine 1) and the subsequent stabilization of the converted structures[1], which would be congruent with such selective[3,9,10] processes. As compelling as this possibility is, however, the data are also consistent with the possibility that new spines formed through instructive processes, for example, by spine extension at particular locations in response to specific cues (Fig. 1b, spine 2).

Of particular interest is the finding that training-associated spine remodelling was not observed in mice that failed to learn the reaching task or could not reach the seeds[1], or in animals exposed to a slowly rotating rod task[2] (which does not require learning). Whereas the latter case can be explained by the failure of new spines to satisfy a certain short-term functionality criterion (such as causally linked pre- and postsynaptic activities[11]), the former can only be explained by a behaviourally driven spine formation and stabilization criterion, operating at longer timescales (hours and days, rather than minutes). The most likely criterion would be a reward-related one[12] — spines would be stabilized only if their functionality resulted in rewarded actions.

This strongly implicates the involvement of diffuse modulatory brain systems — small groups of neurons located in specific brain regions (mainly the brainstem), whose projections cover large brain regions[13]. Most notable in this respect is the dopaminergic system, whose activation has been shown time and again to have key roles in various forms of reinforcement, or reward-driven learning. Indeed, it was recently reported[14] that the elimination of dopaminergic terminals within the same cerebral cortex regions concerned with forelimb movement specifically impaired the acquisition of a motor skill (reaching for a food pellet, a task similar to that studied by Xu and colleagues[1]) but not the execution of a previously acquired skill. Novelty, another attribute ascribed to these modulatory brain systems, also seemed to be extremely important, as re-exposure of trained animals to the original training tasks did not lead to particular spine remodelling, whereas learning new motor tasks did[1,2].

Given the sparseness of the task-related synapse groups[1,2], it is not surprising that Xu *et al.* found that the overlap between synapse groups related to two different tasks was negligible (Fig. 1c). This finding, the fact that most task-related remodelling involved new spines, and the lack of clear size changes in pre-existing spines following training[1], would seem to be at odds with certain tenets of the cell-assembly hypothesis of memory[11,15]. The cell-assembly hypothesis assumes that what distinguishes one memory trace from another is the composition of cells that are co-activated within a given network. Most notably, this hypothesis posits that these distinct compositions are created during memory formation by retuning the strengths of all the synapses already in the network, along a continuum of synaptic strengths. The new results, however, point to strong associations between learned tasks and specific, non-overlapping groups of novel synapses whose strength might remain relatively constant once formed. These 'synapse assemblies' might thus be viewed as the fundamental engrams for these specific motor tasks. Synapse assemblies in a given network would hold the keys to the functioning of that network; for every learned task, the network would function with a connectivity scheme that had been formed and optimized specifically for that task.

Clearly, further studies are required to determine whether the experience-associated spine remodelling described here can be generalized to other forms of memory and other brain systems. Investigating the involvement of diffuse modulatory systems in the process is another intriguing challenge. And although it remains to be shown conclusively that these forms of spine remodelling are essential components of long-term learning and not merely distant echoes of other, yet to be discovered processes, these exciting studies make a convincing case for a

structural basis to skill learning and reopen the field for new theories of memory formation. ■

Noam E. Ziv is in the Department of Physiology and Biophysics, Technion Faculty of Medicine, and Network Biology Research Laboratories, Lokey Center for Life Sciences & Engineering, Haifa 32000, Israel. Ehud Ahissar is in the Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel.
e-mails: noamz@netvision.net.il; ehud.ahissar@weizmann.ac.il

1. Xu, T. *et al. Nature* **462,** 915–919 (2009).
2. Yang, G., Pan, F. & Gan, W.-B. *Nature* **462,** 920–924 (2009).
3. Holtmaat, A. & Svoboda, K. *Nature Rev. Neurosci.* **10,** 647–658 (2009).
4. Denk, W., Strickler, J. H. & Webb, W. W. *Science* **248,** 73–76 (1990).
5. Bhatt, D. H., Zhang, S. & Gan, W. B. *Annu. Rev. Physiol.* **71,** 261–282 (2009).
6. Alvarez, V. A. & Sabatini, B. L. *Annu. Rev. Neurosci.* **30,** 79–97 (2007).
7. Knott, G. W., Holtmaat, A., Wilbrecht, L., Welker, E. & Svoboda, K. *Nature Neurosci.* **9,** 1117–1124 (2006).
8. Nägerl, U. V., Köstinger, G., Anderson, J. C., Martin, K. A. & Bonhoeffer, T. *J. Neurosci.* **27,** 8149–8156 (2007).
9. Jontes, J. D. & Smith, S. J. *Neuron* **27,** 11–14 (2000).
10. Holtmaat, A., Wilbrecht, L., Knott, G. W., Welker, E. & Svoboda, K. *Nature* **441,** 979–983 (2006).
11. Hebb, D. O. *A Neuropsychological Theory* (Wiley, 1949).
12. Crow, T. J. *Nature* **219,** 736–737 (1968).
13. Kety, S. S. in *Neurosciences: Second Study Program* (ed. Schmitt, F. O.) 324–336 (Rockefeller Univ. Press, 1970).
14. Molina-Luna, K., Pekanovic, A., Röhrich, S., Hertler, B. & Schubring-Giese, M. *PLoS ONE* **4,** e7082 (2009).
15. Hopfield, J. J. *Proc. Natl Acad. Sci. USA* **79,** 2554–2558 (1982).

MICROSCOPY

# Photons and electrons team up

F. Javier García de Abajo

**An imaging technique has been demonstrated that blends the principles of conventional light and electron microscopy. It renders images with nanometre and femtosecond space–time resolution.**

How can we probe light fields in the vicinity of nanostructures? This question is becoming increasingly relevant as the confinement and steering of light at the nanoscale are gathering momentum in applications such as optical sensing and information processing. On page 902 of this issue, Barwick *et al.*[1] provide a practical answer to this question. They report a microscopy technique that probes such fields by focusing both light and electron pulses on the nanostructures under study.

In the new imaging technique, which Barwick and colleagues baptized photon-induced near-field electron microscopy (PINEM), a specially designed electron microscope is used to project magnified images of nanostructures in much the same way as an overhead projector forms images of slides from a light beam passing through them. In PINEM, the beam is made of electrons and the role of the slides in the projector is played by light trapped in the vicinity of the samples (see Fig. 4 on page 905).

The term 'trapped' means that the light waves do not propagate and decay exponentially in intensity with distance from the sample. Unlike freely propagating light, this trapped light, called an evanescent light field, can interact efficiently with the electrons, which gain or lose energy through the absorption or emission of light quanta (photons). By means of energy-filtering, the microscope subsequently selects and collects only those electrons that have undergone energy gain, forming images in a process that retains the sub-nanometre spatial resolution that is characteristic of conventional transmission electron microscopes. The number of collected electrons is proportional to the strength of the evanescent field.

Travelling at 70% of the speed of light, the electrons used by Barwick *et al.*[1] spend only a fraction of a femtosecond (1 femtosecond is $10^{-15}$ seconds) near their 100-nanometre-thick samples (carbon nanotubes or silver nanowires). To observe a sizeable, and useful, electron–light interaction during such a short time interval requires intense light fields. In their experiment, the authors achieved high-intensity fields by using two synchronized femtosecond light pulses: one of the pulses was directed to the microscope's electron gun, which converted it into an electron pulse via photoemission; the other, with a peak intensity of about 10–100 gigawatts per square centimetre, was aimed at the nanostructure and was capable of producing multiple-photon absorption (or emission) events by each passing electron — up to eight photons, as the authors report[1]. By design, both pulses have a similar temporal duration, of the order of a few hundred femtoseconds (this duration defines the time resolution of PINEM), and their relative delay in the time of arrival at the sample was controlled with femtosecond precision through the difference in optical path length between the two original light pulses.

Light–electron interactions similar to those seen in Barwick and co-workers' experiment have been observed previously — for example, when electrons pick up thermal phonons (quanta of atomic lattice vibrations) in insulator films[2] or plasmons (quanta of collective oscillations of conduction electrons) from a metal surface[3], or when electrons interact with evanescent fields reflected from an illuminated diffraction grating[4]. Barwick *et al.* are the first to exploit such interactions to image an evanescent light field in the vicinity of a nanostructure, and they accomplish that with nanometre and femtosecond space–time resolution (see Fig. 2 on page 903).

Evanescent light fields have been probed previously with near-field optical microscopes[5] that rely on scanning a subwavelength-sized tip over the sample to form images. These microscopes can yield femtosecond time resolution but are limited by the size of the tip to tens of nanometres in spatial resolution. In addition, the tip can produce undesired artefacts in the images. By contrast, Barwick and colleagues' PINEM technique achieves the sub-nanometre spatial resolution that electron microscopes do. What's more, the electrons constitute a relatively 'clean' probe: moderate electron-beam intensities cause only marginal perturbations in the sample, thus allowing faithful imaging.

Barwick *et al.* demonstrated their PINEM technique for light pulses lasting about 220 femtoseconds, and observed the real-time evolution of the evanescent field that mimics the light pulses themselves. With 220-femtosecond pulses, one could also investigate the dynamics of optical excitations in the sample that have comparable or larger lifetimes, such as certain long-lived photon states that occur in insulating structures (for example, 'Mie modes' in silicon cavities). However, impressive as it is, the technique needs to be adapted for shorter light pulses that can follow the ultrafast dynamical optical response of many nanostructures of interest, such as metallic nanoparticles whose plasmons typically live for only a few tens of femtoseconds.

The propagation of light fields along the surface of a nanostructure is a key ingredient of nanophotonic devices, which carry and process optical signals[6]. The PINEM technique could be improved to study such propagation by sampling the evanescent decay of such fields along the direction perpendicular to the sample surface. This and other developments will surely show that, with Barwick and colleagues' electrons gaining energy, the scientific community will also gain new means of observing the nanoworld. ■

F. Javier García de Abajo is at the Institute of Optics, CSIC, Serrano 121, 28006 Madrid, Spain.
e-mail: jga@cfmac.csic.es

1. Barwick, B., Flannigan, D. J. & Zewail, A. H. *Nature* **462,** 902–906 (2009).
2. Boersch, H., Geiger, J. & Stickel, W. *Phys. Rev. Lett.* **17,** 379–381 (1966).
3. Schilling, J. & Raether, H. *J. Phys. C* **6,** L358–L360 (1973).
4. Mizuno, K., Pae, J., Nozokido, T. & Furuya, K. *Nature* **328,** 45–47 (1987).
5. Balistreri, M. L. M. *et al. Science* **294,** 1080–1082 (2001).
6. Bozhevolnyi, S. I., Volkov, V. S., Devaux, E., Laluet, J. Y. & Ebbesen, T. W. *Nature* **440,** 508–511 (2006).

OBITUARY

# Claude Lévi-Strauss (1908–2009)

Leading anthropologist of his generation.

Claude Lévi-Strauss had only the slightest experience of ethnographic fieldwork, and had no formal training in anthropology. Nevertheless, his ideas transformed the discipline, and profoundly influenced the other human sciences. He died on 31 October.

The son of an artist father, and the grandson of a rabbi in Strasbourg, France, Lévi-Strauss was educated in law and philosophy at the Sorbonne in Paris. In 1935, he joined a French contingent at the new University of São Paulo in Brazil. He embraced the opportunity, because he had a very particular ambition. Jean-Jacques Rousseau once suggested that an expedition should be sent to the Americas to study human nature in its essential state, uncorrupted by civilization. A devotee of Rousseau's philosophy, Lévi-Strauss was determined to execute the master's plan.

During his university holidays, Lévi-Strauss made expeditions to study remote Amerindian settlements, but almost all of this work was done at telegraph posts where Indians were in contact with government agencies and traders. Only towards the end of his travels did he make brief contact with an isolated band living in the old style, but naturally enough they spoke no Portuguese. "Alas! They were only too savage," Lévi-Strauss reported. "They were as close to me as a reflection in a mirror; I could touch them, but I could not understand them."

This adventure might have been no more than an interlude. But shortly after Lévi-Strauss returned to France, the German army invaded the country and he became an exile in New York. At a loose end, he wrote up his Brazilian field notes, became an informal member of the anthropology circle at Columbia University and spent long days combing the anthropology shelves in the New York Public Library. "What I know of anthropology I learned during those years," he later remarked.

Although his American colleagues were steeped in regional ethnography, they were notoriously suspicious of theoretical abstractions. Lévi-Strauss, however, was determined to use observations of hunter-gatherers as the basis for a theory of human nature, like a more empirical Rousseau. Such a soaring ambition required a new theoretical framework. Inspiration came from a fellow exile in New York, the Russian linguist Roman Jakobson. Following the path-breaking contributions of Ferdinand de Saussure, linguistics was in the throes of a theoretical revolution. Lévi-Strauss concluded that it would show the way to a new, generalizing, structural anthropology.

J. ROBINE/AFP/GETTY IMAGES

Jakobson was particularly interested in phonemics, the branch of linguistics that deals with the communication of meaning through sounds. He claimed to have split the atom of linguistics, the phoneme. The phoneme had been viewed as the smallest significant unit of sound in speech, but according to Jakobson it was itself a bundle of features made up of pairs of contrasting elements. So, for instance, English speakers invest the contrasting b and p sounds with meaning (the words 'bill' and 'pill' are obviously different to our ears), whereas in other languages the distinction may be unmarked and unheard. Lévi-Strauss argued that systems of classification are constructed on a similar pattern of binary oppositions.

Returning to Paris in 1949, Lévi-Strauss found employment at the Museum of Man and then the École Pratique des Hautes Études at the Sorbonne. His massive doctoral thesis, published that year, argued that the imposition of a taboo on incest marked the break between a natural and a cultural order, obliging men to exchange sisters with other men, and so creating family and kinship networks. In the simplest (and implicitly oldest) systems, these networks are structured by a binary classification of relatives into two classes: unmarriageable kin and marriageable affines.

In 1959, Lévi-Strauss was appointed to a chair at the College of France in Paris. He now began to publish his most influential studies, which dealt with systems of thought. *The Savage Mind* (1962) argued that Native Americans, and other hunter-gatherers from Australia to Africa, operate a "logic of the concrete" — they order images taken from the world around them in a series of binary oppositions. This home-made natural science provides metaphors for social relationships. In many languages, for example, the Sun and the Moon are associated with male and female characteristics. Paired species of birds and animals are contrasted in terms of colouring, feeding habits, or some other defining features, and are then associated with other paired objects (Australian Aboriginals, for instance, associate the Sun with the crow and the eaglehawk with the Moon). Such categorizations symbolize the social contrasts between men and women, or between pairs of human clans or occupations.

In his masterpiece — a collection of four volumes on American mythology, beginning with the *The Raw and the Cooked* in 1964 and culminating with *The Naked Man* in 1971 — Lévi-Strauss demonstrated how systems of classification are put to work in myths that are at once epics, moral treatises and accounts of the world. He proposed that the initial premises of myths (say, that women are lunar, men solar) are played with in subsequent versions to yield new premises. These transformations follow implicit rules that allow only a sort of logical progression, in the form of the inversion of the initial terms, or a series of substitutions by which one binary pair replaces another. As he put it: "The kind of logic in mythical thought is as rigorous as that of modern science, and ... the difference lies, not in the quality of the intellectual process, but in the nature of things to which it is applied."

Lévi-Strauss notoriously claimed that he had a neolithic intelligence, that his thought was intuitively sympathetic to that of hunter-gatherers. His grand theory rested on a binary opposition between nature and culture (taken, of course, from Rousseau), and so between the small-scale, technically simple societies that existed during the first 150,000 years of human history and modern civilization, which rendered the simpler societies obsolete. He believed that humanity in its natural condition was adapted to the environment, whereas civilized societies endanger the environment and obliterate cultural variation. This deeply pessimistic view was conditioned by the Amazonian idyll of his youth, and the European catastrophe of the Second World War that followed. But he also believed that people everywhere ultimately thought in the same way, although about different things, and that the clash of cultures is necessary for human adaptation.

**Adam Kuper**
Adam Kuper is visiting professor at the Department of Anthropology, Yale University, New Haven, Connecticut 06520, USA.
e-mail: adam.kuper@googlemail.com

ARTICLES

# Probabilistic assessment of sea level during the last interglacial stage

Robert E. Kopp[1,2], Frederik J. Simons[1], Jerry X. Mitrovica[3], Adam C. Maloof[1] & Michael Oppenheimer[1,2]

**With polar temperatures ~3–5 °C warmer than today, the last interglacial stage (~125 kyr ago) serves as a partial analogue for 1–2 °C global warming scenarios. Geological records from several sites indicate that local sea levels during the last interglacial were higher than today, but because local sea levels differ from global sea level, accurately reconstructing past global sea level requires an integrated analysis of globally distributed data sets. Here we present an extensive compilation of local sea level indicators and a statistical approach for estimating global sea level, local sea levels, ice sheet volumes and their associated uncertainties. We find a 95% probability that global sea level peaked at least 6.6 m higher than today during the last interglacial; it is likely (67% probability) to have exceeded 8.0 m but is unlikely (33% probability) to have exceeded 9.4 m. When global sea level was close to its current level ($\geq -10$ m), the millennial average rate of global sea level rise is very likely to have exceeded 5.6 m kyr$^{-1}$ but is unlikely to have exceeded 9.2 m kyr$^{-1}$. Our analysis extends previous last interglacial sea level studies by integrating literature observations within a probabilistic framework that accounts for the physics of sea level change. The results highlight the long-term vulnerability of ice sheets to even relatively low levels of sustained global warming.**

As a result of industrial activity, greenhouse gas concentrations now exceed levels reached on Earth at any time within the past 800 kyr (ref. 1). Given a climate sensitivity of 2–4.5 °C per doubling of carbon dioxide levels[2], current greenhouse gas concentrations—without considering any further increases—are sufficient to cause an equilibrium warming of 1.4–3.2 °C. Among the many effects expected to accompany this warming is a rise in global sea level (GSL)[2], which is defined as the mean value of local sea level (LSL) taken across the ocean. This rise is driven primarily by thermal expansion of sea water and by melting land ice. Uncertainties in ice sheet behaviour make it difficult to predict sea level rise using prognostic models, but by the end of the twenty-first century, GSL could exceed today's value by more than one metre (refs 3, 4). As changes of this magnitude have no precedent in recorded history, to understand them and to compile observations against which to test models of future climate change, it is necessary to turn to the geological record.

In this Article, we analyse a new compilation of geographically dispersed sea level indicators spanning the last interglacial stage (LIG), which climaxed about 125,000 years ago (125 kyr ago). The LIG (also known as the Eemian stage, its local northern European name, and as Marine Isotope Stage 5e) is of special interest for three reasons: (1) it is recent enough that it is possible to obtain some sea level records with high temporal resolution and many more observations with lower temporal resolution; (2) due in large part to enhanced Northern Hemisphere insolation, global and polar temperatures may have been slightly warmer than at present; and (3) several lines of evidence suggest that GSL was higher than today, perhaps by 4–6 m (ref. 1), and that the Greenland Ice Sheet and possibly also the West Antarctic Ice Sheet[5,31] were significantly smaller than they are now.

During the LIG, greenhouse gas concentrations were comparable to pre-industrial Holocene levels[7], but Earth's orbital eccentricity was more than twice the modern value[8]. Energy balance modelling predicts that, as a consequence, summer temperatures between 132 and 124 kyr ago on all land masses except Antarctica were at least 0.5 °C warmer than today[9], while a more complete climate model indicates summer temperatures 2–4 °C warmer than today in most of the Arctic[6]. Ice core data from both Greenland and Antarctica suggest polar temperatures in both hemispheres of about 3–5 °C warmer than today[1], comparable to the 3–6 °C of Arctic warming that is expected to accompany 1–2 °C of global warming[10]. In Europe, pollen data suggest middle Eemian summer temperatures about 2 °C warmer than present[11]. While the change in global mean temperature is uncertain, sea surface temperatures in the equatorial Pacific[12] and Atlantic[13] were about 2 °C warmer than pre-industrial levels.

Synthesizing geological sea level indicators into a global reconstruction requires accounting for regional variability. Differences between LSL and GSL arise because—contrary to an analogy commonly taught in introductory classes—adding water from melting land ice to the ocean is not like pouring water into a bathtub. Many factors other than the changing volume of water in the ocean modulate the influence of melting ice sheets on LSL. These factors include: the direct gravitational effect of the distribution of ice, water and sediment on the sea surface (or geoid), solid Earth deformation and its associated gravitational signature, perturbations to both the magnitude and orientation of the Earth's rotation vector, and time-varying shoreline geometry[14–16], as well as changes in ocean and atmosphere dynamics[17]. In addition, LSLs are influenced by tectonic uplift and thermal subsidence.

As a consequence of these factors, LSLs at Pacific islands far from the late Pleistocene ice sheets were 1–3 m higher in the middle Holocene than today, even though GSL was essentially unchanged[18]. Similarly, even if GSL was never higher than today, LSLs several metres higher than present could have occurred far from the former Laurentide Ice Sheet (for example, in Australia) early in the LIG, and comparably high LSLs could have occurred closer to the former ice sheet (for example, in the Caribbean) late in the LIG[19]. Without accurate and precise dating of the relevant sea level indicators and an appreciation of the difference between LSL and GSL, such patterns could

[1]Department of Geosciences, [2]Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, New Jersey 08544, USA. [3]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA.

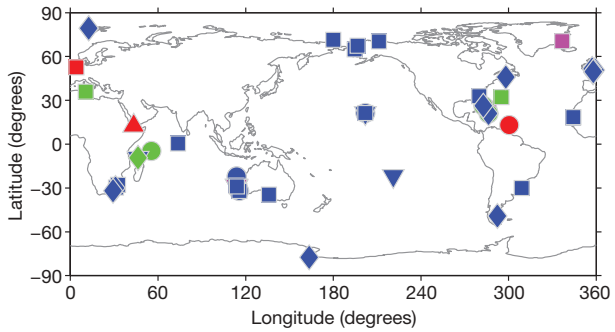



**Figure 1 | Sites with at least one sea level observation in our database.** The symbol shapes reflect the nature of the indicators (upward triangles, isotopic; circles, reef terraces; downward triangles, coral biofacies; squares, sedimentary facies and non-coral biofacies; diamonds, erosional). The colours reflect the number of observations at a site (blue, 1; green, 2; magenta, 3; red, 4 or more).

produce the false appearance of a magnified or diminished GSL highstand. In order to estimate ice sheet history from sea level records, it is thus necessary to account for physical factors like gravitation and solid Earth deformation. Conversely, because these effects cause LSL changes to differ with distance from an ice sheet, a global database of LSL indicators can potentially address not just whether global ice volume was smaller during the LIG than today, but also what combination of melting ice sheets, if any, was responsible for higher GSL.

We construct a database of sea level indicators that is as comprehensive as possible (Figs 1, 2; full data set available in Supplementary Information) and use it to estimate the posterior probability distribution of LSL as a function of space and time and of GSL and ice sheet volumes as functions of time. We must cope with variable geochronological uncertainty, as well as with variable errors in sea levels inferred from proxy data and in estimates of regional long-term tectonic uplift or thermal subsidence. In addition, some of the data provide only upper or lower bounds to sea level. Where possible, we also want to take advantage of quasi-continuous sequences, in which relative timing is known with greater precision than absolute dates. These sequences include a stacked global oxygen isotope curve from benthic foraminifera[20], as well as series of LSL measurements inferred from sedimentary facies in the Netherlands[21] and from hydrological modelling of foraminiferal oxygen isotopes in the Red Sea[22]. (These series are described in detail in Supplementary Information.)

## Statistical approach

The ultimate goal of our analysis is to determine the posterior probability distribution of LIG sea level and ice volume through time, conditioned upon the measurements in our database. Inherent in the method is the assumption that both the prior and posterior distributions are multivariate Gaussian.

We construct a prior probability distribution from the global oxygen isotope curve and its associated age model[20], as described in detail in Methods and Supplementary Information. To do this, we use a physical model of LSL that calculates the eustatic, gravitational, deformational and rotational effects of melting ice sheets[15,16,23]. We estimate the mean and covariance of the prior distribution by averaging the values and covariances of the LSLs and of GSL obtained by running many alternative ice sheet histories through a forward physical model. These histories themselves are sampled from two underlying distributions: a distribution for global ice volume over time based upon ref. 20 and a distribution for individual ice sheet volumes conditioned upon global ice volume. This latter distribution is based upon random perturbations of a model of Last Glacial Maximum (LGM)-to-present ice sheet volume[24] with additional allowances made for ice sheets smaller than their present volumes. To approximate thermosteric effects resulting

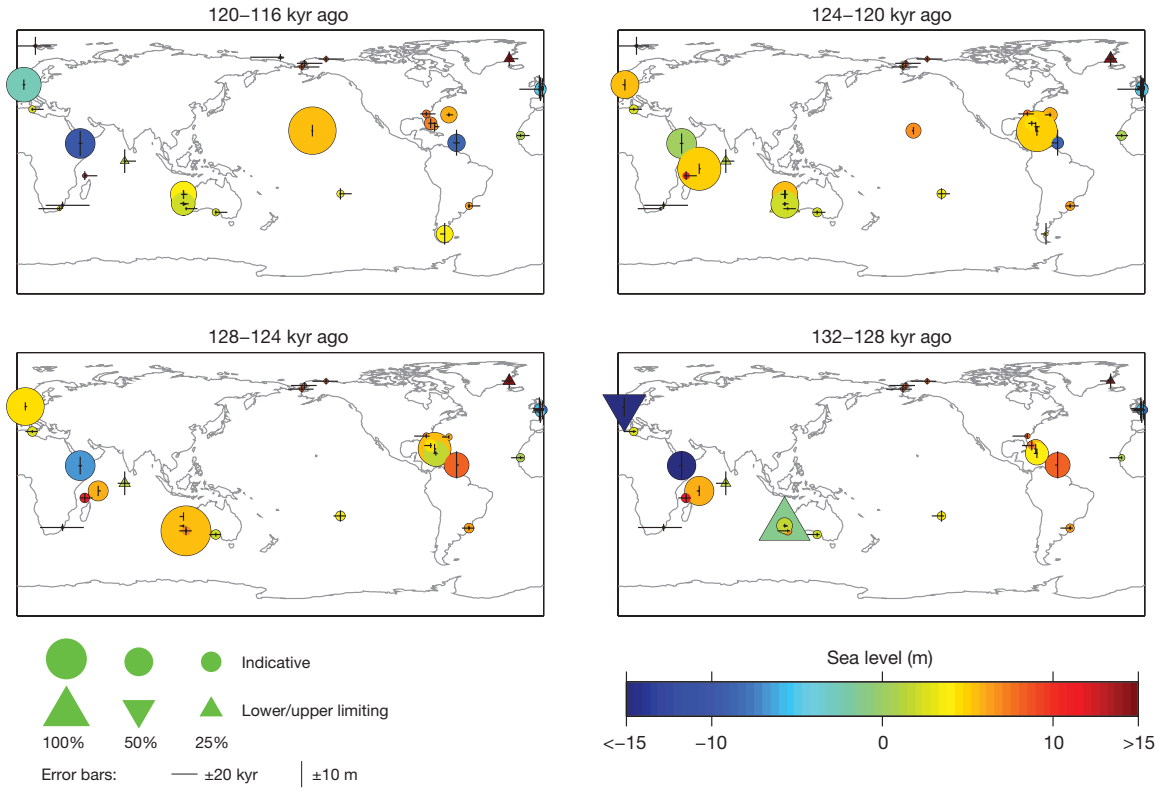



**Figure 2 | Localities at which LSL data exist in our database, for time slices through the LIG.** The diameter of each circle scales as indicated with the probability that the corresponding data point occurs in the indicated interval. The horizontal (vertical) lines are proportional to the standard deviations of the age (sea level) measurements. The intersection of the lines reflects the mean age estimate relative to the age window; a rightward skew reflects a mean estimate earlier than the middle of the window. Data that provide only upper or lower sea level bounds are indicated by downward and upward triangles, respectively. Colours indicate the mean sea level estimate in metres above present value. Some symbols overlap; for a complete table of observations, see Supplementary Information.

from changes in mean ocean temperature and salinity, we add two Gaussian terms: a term independent of time and GSL with a mean of 0 m and a standard deviation of 2 m, and a term that varies with global ice volume ($-1.6 \pm 0.6$ m per 100 m equivalent sea level (e.s.l.) ice sheet growth). The temporal covariance of these thermosteric terms has an e-folding time of 2 kyr. The uncertainty within the thermosteric terms is large enough to also accommodate small contributions from other sources, such as small mountain glaciers present today but not included in the LGM-to-present ice model.

To construct the posterior distribution of sea level at any arbitrary point in space and time, we start with the simpler problem of estimating the posterior probability distribution of sea level at the points included in our database and then interpolate to calculate values at points not in our database. We employ a three-step Gibbs sampler[25] to sample the Bayesian network illustrated in Fig. 3.

In the first step, we calculate corrected measurements of LSLs (**s**) by adjusting the altitude of our proxy observations (**z**) for their depositional settings (**D**), which account for the relationship between proxy altitudes and sea level elevation at the time of formation, and for the background regional uplift or subsidence. The former correction incorporates sedimentological and geomorphological knowledge, such as the fact that most coral observations in the database are of species that grow between 0 and 5 m below mean low tide level[26,27], as well as information about local tidal range. The latter correction is based upon an estimate of the regional uplift or subsidence rate (**u**) and a sample from the posterior distribution of measurement ages (**g**). In selecting or constructing uplift or subsidence rate estimates, we have avoided estimates from the literature that assume LIG sea level as a reference point.

In the second step, we employ Gaussian process regression to estimate the true sea levels (**f**). Gaussian process regression[28], of which the commonly used geospatial technique of kriging interpolation is a well-known example, treats a field (such as sea level) as a collection of random variables drawn from a multivariate Gaussian distribution. By specifying the covariance structure of the field, knowledge about the relevant physics affecting the process can be incorporated into the modelling without constraining it to fit a particular forward model.

In the third step, we use the Metropolis-Hastings algorithm[29] to draw a new Markov chain Monte Carlo sample of the ages (**g**), based upon the measured ages (**t**) and the current estimate of the true sea levels (**f**). Repeating this sequence many times allows us to sample the posterior probability distribution for LSL and GSL in a way that satisfies the measurements to within their uncertainties.

Equipped with an estimate of the posterior probability distribution, we can then answer questions such as 'what was the maximum GSL attained during the LIG' and 'what was the fastest rate at which GSL rose when it was within 10 m of its present value?' (As discussed below, we focus on rates above the $-10$ m threshold because the Laurentide Ice Sheet was comparable in size to the modern Greenland Ice Sheet by the time GSL rose to this level in the Holocene.) To answer such questions, we draw many samples from the posterior distribution and examine the distribution of answers based on these samples. We report these answers as exceedance values. For instance, the 95% probability exceedance value of GSL is exceeded in 95% of all samples. If the 95% exceedance value is 6.6 m, we can reject the hypothesis that sea level never exceeded 6.6 m at the 95% confidence level. Note that the answer to such questions is not identical to the answer one would get by looking at the median projection of GSL and reading its maximum; the maximum of the median would be the 50% probability exceedance value if all time points were perfectly correlated, but such is not the case. The median reconstruction instead represents the best estimate for GSL at each specific point in time, whereas the exceedance values are calculated across the entire LIG interval.

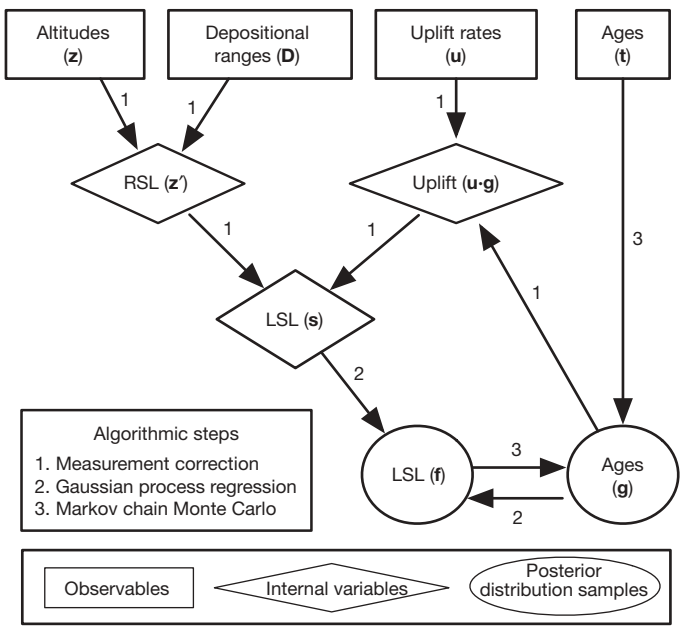



**Figure 3 | Schematic illustration of the process used in our statistical analysis.** See text for details.

## Results of global analysis

Applying our algorithm to the full data set of LIG sea level indicators yields a GSL curve (Fig. 4a) with a median projection that peaks at 124 kyr ago at $7.2 \pm 1.3$ m (67% confidence interval). Further analysis reveals a 95% probability of having exceeded 6.6 m at some time during the LIG highstand and a 67% probability of having exceeded 8.0 m (Fig. 5, solid line). It is unlikely (33% probability) that GSL exceeded 9.4 m.

To test the sensitivity of these results, we analysed seven subsets of the data: one subset excluding the Red Sea oxygen isotope curve, and six either excluding or including only (1) coral data, (2) erosional features, or (3) facies interpretations (Supplementary Information).

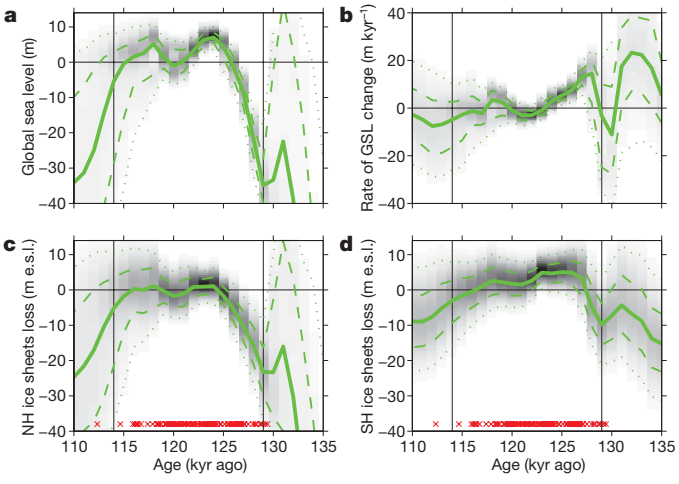



**Figure 4 | Probability density plots of GSL and ice volume during the LIG.** **a**, Global sea level (GSL); **b**, 1,000-year average GSL rates; **c**, Northern Hemisphere (NH) ice volume; and **d**, Southern Hemisphere (SH) ice volume. Heavy lines mark median projections, dashed lines the 16th and 84th percentiles, and dotted lines the 2.5th and 97.5th percentiles. Red crosses mark median posterior estimates of sample ages. Vertical lines mark the interval when >30% of samples from the distribution have standard deviations of GSL <30% of the prior standard deviation (and are thus included in calculations of exceedance probabilities). The horizontal line at 0 indicates modern values in **a**, **c** and **d** and unchanging GSL in **b**. We urge caution in interpreting ice volume projections (**c**, **d**) owing to the use of a Gaussian distribution to represent a non-Gaussian prior. e.s.l., equivalent sea level.

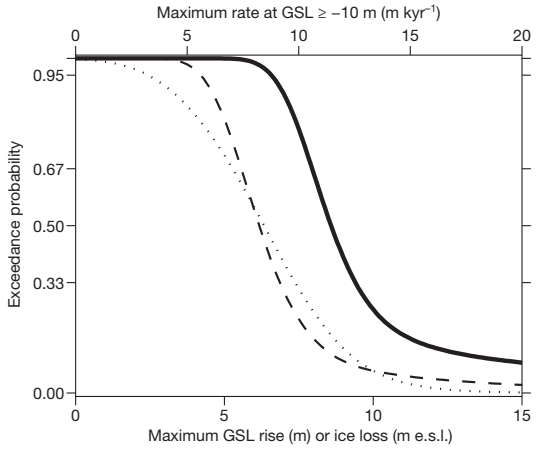



**Figure 5 | Exceedance values calculated from the posterior probability distribution.** The solid line shows GSL rise, the dashed line shows the 1,000-year average rate of change of GSL when GSL is at or above −10 m, and the dotted line shows ice loss in the hemisphere with the least ice loss.

The results from these subsets were fairly consistent. Across all subsets, the median projection peaked between 6.4 and 8.7 m. With the exception of the subset containing only erosional features, the 95% probability exceedance value ranged from 5.7 to 7.0 m, the 67% probability value ranged from 7.3 to 8.7 m, and the 33% probability value ranged from 8.4 to 10.5 m. (The values for the subset containing only erosional features were slightly lower and more broadly spread, with 95%, 67% and 33% values of −0.3 m, 3.9 m and 6.8 m, respectively. The spread reflects the relatively high uncertainty on this projection, which results in large part from a smaller data set.) We therefore consider our results to be reasonably robust with respect to different observations.

The 95%, 67% and 33% probability exceedance values for 1,000-year average GSL rise rate during the interval when GSL was ≥−10 m are 5.6 m kyr$^{-1}$, 7.4 m kyr$^{-1}$ and 9.2 m kyr$^{-1}$, respectively (Fig. 4b; Fig. 5, dashed line). We emphasize that these values by no means exclude faster intervals of sea level rise lasting for less than one millennium.

We can also attempt to answer questions about the magnitude of ice sheet volume based on the posterior probability distribution, but we must do so with caution. The distribution of Northern Hemisphere ice volume, in particular, can only be roughly approximated with a Gaussian, as it has a hard upper bound set by the fact that there is only about 7 m e.s.l. of Northern Hemisphere ice available to melt today. Because of this limitation, although we directly present the hemispheric ice volume posteriors in Fig. 4c, d, we make only one fairly conservative inference regarding ice sheet volumes. The posterior distribution suggests a 95% probability that both Northern Hemisphere ice sheets and Southern Hemisphere ice sheets reached minima at which they were at least 2.5 m e.s.l. smaller than today, although not necessarily at the same point in time (Fig. 5, dotted line). We can make no strong statements about in which hemisphere the ice shrunk to a greater extent; in 59% of samples, it was the Southern Hemisphere and in 41% of samples, it was the Northern Hemisphere. Additional sea level proxies close to the ice sheets would help increase the precision of these estimates, as might a non-Gaussian model for the prior distribution.

## Comparison to previous estimates

Previous estimates of LIG sea level, which were generally in the range 4–6 m, were based on interpretations of LSL at a small number of localities. The Fourth Assessment Report of the IPCC[1] highlighted Hawaii and Bermuda[30]; other authors[31] also include observations from the Bahamas, Western Australia and the Seychelles Islands. All these localities are relatively tectonically stable and experience only slow thermal subsidence, associated with the cooling of the lithosphere. If one had to draw conclusions about GSL from a small number of LSL measurements, these are reasonable sites at which to look.

Other commonly considered localities, such as Barbados[32] and the Huon Peninsula[33], are rapidly uplifting localities. These sites have advantages as relative sea level recorders, most notably that terraces recording sea levels below present are readily accessible. Assuming these sites have experienced a steady rate of uplift, they can help uncover sea level variations over fairly short timescales. However, they are poor sites from which to draw conclusions about absolute sea levels, as recovering this information requires a precise estimate of uplift rate. Because our method incorporates knowledge about the associated uncertainties, we can include both stable and uplifting sites into our analysis.

To our knowledge, only one previous study[19], which used a fairly limited set of observations, has attempted to account for the effects of glacial isostatic adjustment in drawing conclusions about GSL and ice volume from LIG sea level records. As that study demonstrated, understanding the influence of these effects is critical, as otherwise LSL highstands could easily be falsely interpreted as reflecting global highstands. Our statistical model uses the covariance between local and GSL, derived from many runs of a forward physical model, to account for the gravitational, deformational and rotational effects of the ice–ocean mass redistribution. Our results indicate that the apparent high GSL during the LIG is indeed real, though previously underestimated.

## Rates of sea level change

Our results suggest that during the interval of the LIG when sea level was above −10 m, the rate of sea level rise, averaged over 1 kyr, was very likely to have reached values of at least about 5.6 m kyr$^{-1}$ but was unlikely to have exceeded 9.2 m kyr$^{-1}$. Our data do not permit us to resolve confidently rates of sea level change over shorter periods of time. Our inferences are consistent with estimates of the rate of the contribution of Laurentide Ice Sheet meltwater to GSL during the early Holocene; the Laurentide Ice Sheet contribution is estimated to account for about 7 m kyr$^{-1}$ during the period when GSL climbed above −10 m (ref. 34).

Ice volume during the late deglacial rise at the start of the LIG was only slightly larger than at present. The Laurentide Ice Sheet would have been a shrunken remnant of its once extensive mass—or, perhaps two small remnants, one over Québec and Labrador and one over eastern Nunavut and Baffin Island, as in the early Holocene[34,35]. As the Laurentide Ice Sheet was within a factor of two in size of the present Greenland Ice Sheet, its dynamics may have been analogous to those of the Greenland Ice Sheet. The results from the LIG suggest that, given a sufficient forcing, the present ice sheets could sustain a rate of GSL rise of about 56–92 cm per century for several centuries, with these rates potentially spiking to higher values for shorter periods.

## Discussion

Although it is the approach most commonly taken when the LIG is used as an analogue for near-future warming, GSL and global ice volume cannot be accurately inferred by a qualitative examination of LSL at a handful of localities. Better control is afforded by a more thorough approach that combines, as we do, an extensive database of sea level indicators with a probabilistic assessment of their interpretive and geochronological errors. The results of our analysis support the common hypothesis that LIG GSL was above the current value, but contrary to previous estimates, we conclude that peak GSL was very likely to have exceeded 6.6 m and was likely to have been above 8.0 m, though it is unlikely to have exceeded 9.4 m.

The LIG was only slightly warmer than present, with polar temperatures similar to those expected under a low-end, ~2 °C warming scenario. Nonetheless, it appears to have been associated with substantially smaller ice sheets than exist at present. Achieving GSL in excess of 6.6 m higher than present is likely to have required major melting of both the Greenland and the West Antarctic ice sheets, an

inference supported by our finding that both Northern and Southern hemisphere ice volumes are very likely to have shrunk by at least 2.5 m e.s.l. relative to today. Incorporating a large database of palaeoclimatic constraints thus highlights the vulnerability of ice sheets to even relatively low levels of sustained global warming.

## METHODS SUMMARY

We assembled our database, which includes observations from 42 localities, through an extensive literature search for indicators with best estimates of ages between 140 and 90 kyr ago. To each indicator we assigned a depth range of formation or deposition based upon geomorphological and sedimentological interpretation. See Methods and Supplementary Information for full details of the database, the statistical analysis algorithm, and the physical model used to generate the covariance function.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 27 February; accepted 11 November 2009.**


1. Jansen, E. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.* ) 433–498 (Cambridge Univ. Press, 2007).
2. Meehl, G. A. *et al.* in *Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) 747–845 (Cambridge Univ. Press, 2007).
3. Rahmstorf, S. A semi-empirical approach to projecting future sea-level rise. *Science* **315,** 368–370 (2007).
4. Grinsted, A., Moore, J. C. & Jevrejeva, S. Reconstructing sea level from paleo and projected temperatures 200 to 2100 AD. *Clim. Dyn.* doi:10.1007/s00382-008-0507-2 (published online 6 January 2009).
5. Cuffey, K. M. & Marshall, S. J. Substantial contribution to sea-level rise during the last interglacial from the Greenland ice sheet. *Nature* **404,** 591–594 (2000).
6. Otto-Bliesner, B., Marshall, S., Overpeck, J., Miller, G. & Hu, A. Simulating Arctic climate warmth and icefield retreat in the Last Interglaciation. *Science* **311,** 1751–1753 (2006).
7. Petit, J. *et al.* Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399,** 429–436 (1999).
8. Berger, A. & Loutre, M. F. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10,** 297–317 (1991).
9. Crowley, T. & Kim, K. Milankovitch forcing of the Last Interglacial sea level. *Science* **265,** 1566–1568 (1994).
10. Katsov, V. M. *et al.* in *Arctic Climate Impact Assessment* (eds Symon, C., Arris, L. & Heal, B.) Ch. 4, 99–150 (Cambridge Univ. Press, 2004).
11. Kaspar, F., Kühl, N., Cubasch, U. & Litt, T. A model-data comparison of European temperatures in the Eemian interglacial. *Geophys. Res. Lett.* **32,** L11703, doi:10.1029/2005GL022456 (2005).
12. Lea, D. The 100,000-yr cycle in tropical SST, greenhouse forcing, and climate sensitivity. *J. Clim.* **17,** 2170–2179 (2004).
13. Weldeab, S., Lea, D., Schneider, R. & Andersen, N. 155,000 years of West African monsoon and ocean thermal evolution. *Science* **316,** 1303–1307 (2007).
14. Farrell, W. E. & Clark, J. A. On postglacial sea level. *Geophys. J. R. Astron. Soc.* **46,** 647–667 (1976).
15. Mitrovica, J. X. & Milne, G. A. On post-glacial sea level: I. General theory. *Geophys. J. Int.* **154,** 253–267 (2003).
16. Kendall, R., Mitrovica, J. & Milne, G. On post-glacial sea level – II. Numerical formulation and comparative results on spherically symmetric models. *Geophys. J. Int.* **161,** 679–706 (2005).
17. Yin, J., Schlesinger, M. E. & Stouffer, R. J. Model projections of rapid sea-level rise on the northeast coast of the United States. *Nature Geosci.* **2,** 262–266 (2009).
18. Mitrovica, J. X. & Milne, G. A. On the origin of late Holocene sea-level highstands within equatorial ocean basins. *Quat. Sci. Rev.* **21,** 2179–2190 (2002).
19. Lambeck, K. & Nakada, M. Constraints on the age and duration of the last interglacial period and on sea-level variations. *Nature* **357,** 125–128 (1992).
20. Lisiecki, L. E. & Raymo, M. E. A. Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}$O records. *Paleoceanography* **20,** 1–17 (2005).
21. Zagwijn, W. H. Sea-level changes in the Netherlands during the Eemian. *Geol. Mijnb.* **62,** 437–450 (1983).
22. Rohling, E. J. *et al.* High rates of sea-level rise during the last interglacial period. *Nature Geosci.* **1,** 38–42 (2008).
23. Mitrovica, J., Wahr, J., Matsuyama, I. & Paulson, A. The rotational stability of an ice-age earth. *Geophys. J. Int.* **161,** 491–506 (2005).
24. Peltier, W. R. Global glacial isostasy and the surface of the ice-age Earth: the ICE-5G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.* **32,** 111–149 (2004).
25. Banerjee, S., Carlin, B. P. & Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data* (Chapman & Hall/CRC, 2003).
26. Lighty, R. G., Macintyre, I. G. & Stuckenrath, R. *Acropora palmata* reef framework: a reliable indicator of sea level in the western Atlantic for the past 10,000 years. *Coral Reefs* **1,** 125–130 (1982).
27. Camoin, G. F., Ebren, P., Eisenhauer, A., Bard, E. & Faure, G. A 300,000-yr coral reef record of sea level changes, Mururoa atoll (Tuamotu archipelago, French Polynesia). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **175,** 325–341 (2001).
28. Rasmussen, C. & Williams, C. *Gaussian Processes for Machine Learning* (MIT Press, 2006).
29. Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109 (1970).
30. Muhs, D. R., Simmons, K. R. & Steinke, B. Timing and warmth of the Last Interglacial period: new U-series evidence from Hawaii and Bermuda and a new fossil compilation for North America. *Quat. Sci. Rev.* **21,** 1355–1383 (2002).
31. Overpeck, J. T. *et al.* Paleoclimatic evidence for future ice-sheet instability and rapid sea-level rise. *Science* **311,** 1747–1750 (2006).
32. Schellmann, G. & Radtke, U. A revised morpho- and chronostratigraphy of the Late and Middle Pleistocene coral reef terraces on Southern Barbados (West Indies). *Earth Sci. Rev.* **64,** 157–187 (2004).
33. Esat, T. M., McCulloch, M. T., Chappell, J., Pillans, B. & Omura, A. Rapid fluctuations in sea level recorded at Huon Peninsula during the penultimate glaciation. *Science* **283,** 197–202 (1999).
34. Carlson, A. E. *et al.* Rapid early Holocene deglaciation of the Laurentide ice sheet. *Nature Geosci.* **1,** 620–624 (2008).
35. Tamisiea, M. E., Mitrovica, J. X. & Davis, J. L. GRACE gravity data constrain ancient ice geometries and continental dynamics over Laurentia. *Science* **316,** 881–883 (2007).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank J. R. Stroud, L. D. Brown and B. McShane for statistical guidance, and M. Bender, B. Horton, D. Nychka and D. Peltier for comments. We also thank G. Spada for providing his SELEN sea level code, which we used for preliminary calculations incorporated in a previous version of the statistical model. Computing resources were substantially provided by the TIGRESS high performance computer centre at Princeton University, which is jointly supported by the Princeton Institute for Computational Science and Engineering and the Princeton University Office of Information Technology. R.E.K. was supported by a postdoctoral fellowship in the programme on Science, Technology, and Environmental Policy at the Woodrow Wilson School of Princeton University.

**Author Contributions** R.E.K. compiled the database, developed the statistical analysis method, and co-wrote the paper. F.J.S. contributed to the development of the statistical analysis method and co-wrote the paper. J.X.M. developed the physical sea level model and co-wrote the paper. A.C.M. contributed to the compilation of the database and co-wrote the paper. M.O. supervised the project and co-wrote the paper.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.E.K. (rkopp@alumni.caltech.edu).

## METHODS

**Database of LIG sea level indicators.** We characterize each LIG sea level indicator (indexed by $i$) by five parameters: its geographical position ($\mathbf{r}_i$), its measured altitude with respect to mean tide level ($z_i$), its measured age ($t_i$), the range of depths at which it might have formed ($\mathbf{D}_i$), and the estimated local uplift or subsidence rate ($u_i$). Some of the observations are censored, in that they provide only an upper or lower bound to sea level. When more than one observation comes from the same locality, we also record stratigraphic order and, where available, estimates of the relative ages of observations. With the exception of geographical position, each of these variables has uncertainties that we assume follow a Gaussian distribution. For some values, including all depositional depth ranges, uniform distributions between two limits $a$ and $b$ may be a better choice than Gaussian ones. In these cases, we substitute a Gaussian distribution with the same mean and standard deviation as the uniform distribution, that is, $(b-a)/\sqrt{12}$. Depositional ranges $\mathbf{D}_i$ are thus replaced with Gaussian estimates $d_i$. The full database is supplied in Supplementary Information.

**Prior distribution.** We assume that sea level is a Gaussian process with a spatially and temporally varying covariance described by the function $k(\mathbf{r}_i, g_i; \mathbf{r}_j, g_j)$. There is no uncertainty on spatial location $\mathbf{r}_i$, but the temporal variable is $g_i$, the model age (see Fig. 3). We approximate $k$ by $\hat{k}$, which is produced by sampling alternative histories from a forward model that incorporates the relevant physics. To stabilize the estimate and reduce variability related to finite sample size, we smooth $\hat{k}$ with a Gaussian temporal taper function: $\hat{k}(\mathbf{r}_i, g_i\,;\, \mathbf{r}_j, g_j) = \hat{k}_0(\mathbf{r}_i, g_i;\, \mathbf{r}_j, g_j)\exp\left[-(g_i-g_j)^2/\tau^2\right]$, as discussed in the Supplementary Information. To produce the results described in the main text, we employed $\tau = 3$ kyr. Results from other values are shown in Supplementary Information.

The prior probability distribution is based upon the age model of ref. 20, which places the start of the deglaciation at about 135 kyr ago and the start of the LIG highstand at about 127 kyr ago. For consistency, we have aligned the Red Sea and Dutch sequences against this record and excluded from the main analysis three observations from the Houtman-Abrohlos Islands[36,37] whose ages are inconsistent with this model. There is, however, considerable disagreement among current age models. Reference 38 (adopted in ref. 22) places the start of the highstand at about 125 kyr ago, 2 kyr later than ref. 20, while ref. 39 places the start of the deglaciation at between 137 and 142 kyr ago, 2–7 kyr earlier. Our results do not attempt to address these differences, and should be viewed in the context of the ref. 20 timescale.

**Physical model.** The physical model is based on a gravitationally self-consistent sea-level equation[15] that extends earlier work[14] to take exact account of shoreline migration due to either local sea-level changes (which give rise to offlap or onlap) and changes in the extent of grounded, marine-based ice. The calculations are performed using a pseudo-spectral sea-level solver[16,40] with a truncation at spherical harmonic degree and order 256. The solver incorporates the feedback on sea level of contemporaneous, load-induced perturbations in the Earth's rotation vector[16], where these perturbations are computed using the new ice-age rotation theory of ref. 23. The sensitivity to Earth structure is embedded within viscoelastic surface load and tidal Love numbers[41,42]. We adopt spherically symmetric, self-gravitating, Maxwell viscoelastic Earth models. The elastic and density structure of these models is given by the seismic model PREM (ref. 43). The viscosity profile is discretized into three layers, including: (1) an extremely high (essentially elastic) lithospheric lid of thickness LT; (2) a uniform viscosity from the base of the lithosphere to 670 km depth (that is, the sub-lithospheric upper mantle) which we denote as $\nu_{\mathrm{UM}}$; and (3) a uniform lower mantle viscosity (that is, from 670 km depth to the core-mantle boundary) denoted by $\nu_{\mathrm{LM}}$. We consider a suite of 72 such Earth models generated by using the following choices: LT = 70, 95, or 120 km; $\nu_{\mathrm{UM}} = 0.3$, 0.5, 0.8 or $1.0 \times 10^{21}$ Pa s; $\nu_{\mathrm{LM}} = 2$, 3, 5, 8, 10, or $20 \times 10^{21}$ Pa s.

As described in Supplementary Information, we generate an estimate of the prior sea level covariance $\hat{k}$ by running the model 250 times with different ice sheet histories and randomly selected viscosity profiles. From these runs, we compute the covariance among LSLs at evenly spaced points, GSLs and ice sheet volumes, as well as at the exact coordinates of the sites in our database, and we store the results as a lookup table. Total ice volume in the different ice sheet histories is sampled from a distribution based upon the ref. 20 global oxygen isotope curve. The ice volume of individual ice sheets is sampled from a probability distribution for individual ice sheet volumes that is conditional upon total global ice volume. This latter distribution is constructed from random perturbations of LGM-to-present ice models[24].


36. Eisenhauer, A., Zhu, Z., Collins, L., Wyrwoll, K. & Eichstätter, R. The Last Interglacial sea level change: new evidence from the Abrolhos islands, West Australia. *Int. J. Earth Sci.* **85,** 606–614 (1996).
37. Zhu, Z. R. *et al.* High-precision U-series dating of Last Interglacial events by mass spectrometry: Houtman Abrolhos Islands, western Australia. *Earth Planet. Sci. Lett.* **118,** 281–293 (1993).
38. Thompson, W. G. & Goldstein, S. L. Open-system coral ages reveal persistent suborbital sea-level cycles. *Science* **308,** 401–405 (2005).
39. Thomas, A. L. *et al.* Penultimate deglacial sea-level timing from uranium/thorium dating of Tahitian corals. *Science* **324,** 1186–1189 (2009).
40. Mitrovica, J. X. & Peltier, W. R. On postglacial geoid subsidence over the equatorial ocean. *J. Geophys. Res.* **96,** 20053–20071 (1991).
41. Peltier, W. R. The impulse response of a Maxwell Earth. *Rev. Geophys. Space Phys.* **12,** 649–669 (1974).
42. Wu, P. *The Response of a Maxwell Earth to Applied Surface Mass Loads: Glacial Isostatic Adjustment.* M.Sc. thesis, Univ. Toronto (1978).
43. Dziewonski, A. & Anderson, D. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25,** 297–356 (1981).

Vol 462|17 December 2009|doi:10.1038/nature08625

ARTICLES

# Parental origin of sequence variants associated with complex diseases

Augustine Kong[1], Valgerdur Steinthorsdottir[1]*, Gisli Masson[1]*, Gudmar Thorleifsson[1]*, Patrick Sulem[1], Soren Besenbacher[1], Aslaug Jonasdottir[1], Asgeir Sigurdsson[1], Kari Th. Kristinsson[1], Adalbjorg Jonasdottir[1], Michael L. Frigge[1], Arnaldur Gylfason[1], Pall I. Olason[1], Sigurjon A. Gudjonsson[1], Sverrir Sverrisson[1], Simon N. Stacey[1], Bardur Sigurgeirsson[2], Kristrun R. Benediktsdottir[3], Helgi Sigurdsson[4], Thorvaldur Jonsson[5], Rafn Benediktsson[6], Jon H. Olafsson[2], Oskar Th. Johannsson[4], Astradur B. Hreidarsson[6], Gunnar Sigurdsson[6], the DIAGRAM Consortium†, Anne C. Ferguson-Smith[7], Daniel F. Gudbjartsson[1], Unnur Thorsteinsdottir[1,8] & Kari Stefansson[1,8]

**Effects of susceptibility variants may depend on from which parent they are inherited. Although many associations between sequence variants and human traits have been discovered through genome-wide associations, the impact of parental origin has largely been ignored. Here we show that for 38,167 Icelanders genotyped using single nucleotide polymorphism (SNP) chips, the parental origin of most alleles can be determined. For this we used a combination of genealogy and long-range phasing. We then focused on SNPs that associate with diseases and are within 500 kilobases of known imprinted genes. Seven independent SNP associations were examined. Five—one with breast cancer, one with basal-cell carcinoma and three with type 2 diabetes—have parental-origin-specific associations. These variants are located in two genomic regions, 11p15 and 7q32, each harbouring a cluster of imprinted genes. Furthermore, we observed a novel association between the SNP rs2334499 at 11p15 and type 2 diabetes. Here the allele that confers risk when paternally inherited is protective when maternally transmitted. We identified a differentially methylated CTCF-binding site at 11p15 and demonstrated correlation of rs2334499 with decreased methylation of that site.**

The effect of sequence variants on phenotypes may depend on parental origin. The most obvious scheme, although not the only one[1], is imprinting in which the effect is limited to the allele inherited from a parent of a specific sex. Despite this, most reports of genome-wide association studies have treated the paternal and maternal alleles as exchangeable. This is understandable, as the information required is often unavailable, but it reduces the power of such studies to discover some susceptibility variants and underestimates the effects of others, contributing to unexplained heritability. Here we describe a method that allows us to determine the parental origin of haplotypes systematically even when the parents of probands are not genotyped. We use the results to discover associations that exhibit parental-origin-specific effects.

## Determining parental origin

Long-range phasing allows for accurate phasing of Icelandic samples typed with Illumina BeadChips for regions up to 10 cM in length[2]. Two advances have been made since then, stitching and parental-origin determination. Genome-wide, long-range phasing was applied to overlapping tiles, each 6 cM in length, with 3-cM overlaps between consecutive tiles. For each tile, we attempted to determine the parental origins of the two phased haplotypes regardless of whether the parents of the proband were chip-typed. Using the Icelandic genealogy database, for each of the two haplotypes of a proband a search was performed to identify, among those individuals also known to carry the same haplotype, the closest relative on each of the paternal and maternal sides (Fig. 1). Results for the two haplotypes were combined into a robust single-tile score reflecting the relative likelihood of the two possible parental-origin assignments

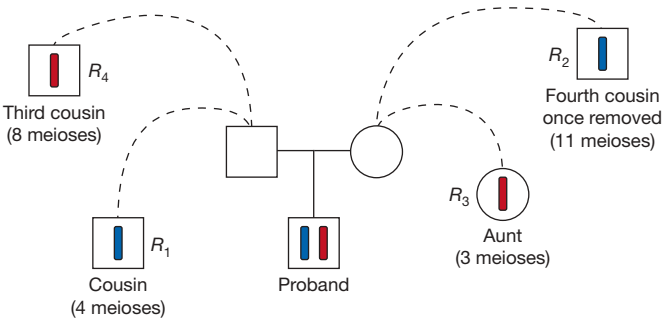



**Figure 1 | An example of determination of parental origin.** In blue and red are two phased haplotypes of a proband. Among other typed individuals, the closest paternal relative known also to carry the blue haplotype is $R_1$, a cousin; the corresponding maternal relative is $R_2$. For the red haplotype, a maternal aunt ($R_3$) carries the haplotype, and the closest known carrier on the paternal side is $R_4$. Because $R_1$ is a closer relative than $R_2$, and $R_3$ is a closer relative than $R_4$, the blue and red haplotypes are probably paternally and maternally inherited, respectively. The single-tile score (Methods) supporting this assignment is 0.194.

[1]deCODE genetics, Sturlugata 8, 101 Reykjavík, Iceland. [2]Department of Dermatology, [3]Department of Pathology, [4]Department of Oncology, [5]Department of Surgery, [6]Department of Endocrinology and Metabolism, Landspitali-University Hospital, 101 Reykjavik, Iceland. [7]Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 3EG, UK. [8]Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.
*These authors contributed equally to this work.
†Lists of participants and affiliations appear at the end of the paper.

(with a score greater than zero supporting one assignment and a score less than zero supporting the other assignment; see Methods for details). We then tried to stitch the haplotypes from consecutive tiles together on the basis of sharing at the overlapping region. Stitching and parental-origin determination are complementary tasks. Specifically, if parental origin is determined with high confidence for one tile, the information can be propagated to other tiles through stitching (Supplementary Fig. 1a). Conversely, in cases in which the overlap between two adjacent tiles is homozygous for all SNPs, stitching can still be accomplished if parental origins can be determined for both tiles independently (Supplementary Fig. 1b). For haplotypes derived by stitching, a contig score for parental origin is computed by summing the individual single-tile scores.

After filtering based on various quality and yield criteria, 289,658 autosomal markers and 8,411 markers on chromosome X were used. Excluding those with no parent listed in the genealogy database or with a genotyping yield of less than 98%, 38,167 individuals, the majority typed with Illumina HumanHap300 or CNV370 BeadChips (Supplementary Information), were processed. For these individuals, 97.8% of the heterozygous genotypes were long-range phased, and in 99.8% of these the parental origin was determined. Overall, 3,841,331,873 heterozygous genotypes, or 97.7% of all heterozygous genotypes, had parental origin assigned. The data includes 2,879 typed trios. To evaluate the accuracy of our method empirically, a run was performed with the data for parents in these trios removed when determining parental origin. For 231,585,437 heterozygous genotypes in the probands/offspring, parental origin was determined both by our method and using the trio data directly, with 500,330 discrepancies, an error rate of 0.22%. Because the trios tested passed heritability checks in preprocessing, the error rate for individuals with fewer than two parents genotyped is probably higher. Nevertheless, the overall error rate is probably less than 0.4% (Supplementary Information).

## Imprinting and disease association

Although many mechanisms can lead to parental-origin-specific association with a phenotype, sequence variants located close to imprinted genes are more likely to exhibit such behaviour a priori. Through two sources, ref. 3 and the Imprinted Gene Catalogue[4,5], we found forty-eight genes known to be imprinted in humans (Supplementary Table 1). Selecting regions that fall within 500 kilobases (kb) of any of these genes (NCBI build 36 of the human genome assembly) amounts to approximately 1% of the genome. The 500-kb threshold was chosen because imprinted genes often occur in clusters and the imprinting status of genes close to known imprinted genes is often undetermined. It is also known that a sequence variant can directly affect the function of a gene located some distance away. Among the 298,069 SNPs we processed, 3,840 fall within these selected regions.

By consulting the US National Institutes of Health Office of Population Studies catalogue of published genome-wide association studies[6] (accessed 25 April 2009), we intersected reported SNP–disease associations with $P < 5 \times 10^{-8}$ with the selected regions (Supplementary Table 2). After further restriction to diseases for which genome scans have been published based on Icelandic data, four associations remained. Three other SNP associations we were aware of that fall within the imprinted regions, one recently published for basal-cell carcinoma[7] and two new type 2 diabetes (T2D) variants discovered in the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) sample set (unpublished data, DIAGRAM Consortium; Supplementary Information), were also examined.

## Association analysis

For each disease–SNP association, five tests were performed (Table 1). We performed a standard case-control test without taking parental origin into account to provide a baseline. Then we performed a

**Table 1 | Parental-origin-specific analyses of disease-susceptibility variants**

| Disease, SNP [alleles]* | | Standard case-control test | | Tests of association with parental origins | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| NCBI build 36 position, *N* | *M*, $F_{con}$ | OR | $P$‡ | Paternal allele§ | | Maternal allele§ | | 2-d.f. test‖ | Paternal vs maternal (case only) | |
| | | | | OR | *P* | OR | *P* | *P* | n12:n21¶ | *P* |
| **Breast cancer, rs3817198† [C/T]** | | | | | | | | | | |
| C11 1,865,582, | 34,909, | | | | | | | | | |
| 1,803 | 0.303 | 1.04 | 0.36 | 1.17 | 0.038 | 0.91 | 0.11 | 0.0040 | 437:339 | $6.2 \times 10^{-4}$ |
| **Basal-cell carcinoma, rs157935 [T/G]** | | | | | | | | | | |
| C7 130,236,093, | 37,041, | | | | | | | | | |
| 1,118 | 0.676 | 1.23 | $1.8 \times 10^{-5}$ | 1.40 | $1.5 \times 10^{-6}$ | 1.09 | 0.19 | $3.8 \times 10^{-6}$ | 237:182 | 0.010 |
| **T2D, rs2237892 [C/T]** | | | | | | | | | | |
| C11 2,796,327, | 34,706, | | | | | | | | | |
| 1,468 (discovery) | 0.925 | 1.19 | 0.044 | 1.14 | 0.24 | 1.24 | 0.071 | 0.095 | 81:90 | 0.51 |
| 783 (replication) | | 1.08 | 0.43 | 0.87 | 0.30 | 1.43 | 0.024 | 0.050 | 35:59 | 0.014 |
| 2,251 (combined) | | 1.15 | 0.043 | 1.03 | 0.71 | 1.30 | 0.0084 | 0.027 | 116:149 | 0.054 |
| **T2D, rs231362† [C/T]** | | | | | | | | | | |
| C11 2,648,047, | 33,377, | | | | | | | | | |
| 1,423 (discovery) | 0.551 | 1.09 | 0.051 | 0.97 | 0.67 | 1.23 | 0.0010 | 0.0037 | 329:401 | 0.014 |
| 750 (replication) | | 1.10 | 0.073 | 1.00 | 0.99 | 1.22 | 0.011 | 0.037 | 158:191 | 0.098 |
| 2,173 (combined) | | 1.10 | 0.013 | 0.98 | 0.73 | 1.23 | $6.2 \times 10^{-5}$ | $2.6 \times 10^{-4}$ | 487:592 | 0.0032 |
| **T2D, rs4731702 [C/T]** | | | | | | | | | | |
| C7 130,083,924, | 34,706, | | | | | | | | | |
| 1,468 (discovery) | 0.439 | 1.15 | 0.0018 | 1.07 | 0.24 | 1.23 | $6.4 \times 10^{-4}$ | 0.0013 | 335:374 | 0.17 |
| 783 (replication) | | 0.95 | 0.38 | 0.84 | 0.024 | 1.08 | 0.31 | 0.048 | 163:204 | 0.037 |
| 2,251 (combined) | | 1.08 | 0.039 | 0.99 | 0.79 | 1.17 | 0.0010 | 0.0041 | 498:578 | 0.022 |
| **T2D, rs2334499 [T/C]** | | | | | | | | | | |
| C11 1,653,425, | 34,706, | | | | | | | | | |
| 1,468 (discovery) | 0.412 | 1.11 | 0.017 | 1.41 | $4.3 \times 10^{-9}$ | 0.87 | 0.020 | $3.5 \times 10^{-9}$ | 437:276 | $7.0 \times 10^{-9}$ |
| 783 (replication) | | 1.02 | 0.71 | 1.23 | 0.0055 | 0.84 | 0.023 | 0.0018 | 222:157 | $8.0 \times 10^{-4}$ |
| 2,251 (combined) | | 1.08 | 0.034 | 1.35 | $4.7 \times 10^{-10}$ | 0.86 | 0.0020 | $5.7 \times 10^{-11}$ | 659:433 | $4.1 \times 10^{-11}$ |

NCBI build 36 position is shown in terms of chromosome and base number. *N*, case sample size; *M*, control set size; $F_{con}$, control frequency (frequency of the risk allele in controls).
* The first allele is the risk allele on the basis of analyses that do not take into account parent of origin.
† Imputed allele probabilities were used.
‡ Genomic control was applied (true for all *P* values shown).
§ The effect of the paternally inherited allele was tested by comparing the corresponding alleles in cases with those in controls. The effect of the maternally inherited allele was tested similarly.
‖ The test assumes a multiplicative effect for the paternally and maternally inherited alleles, but allows the effects to be different under the alternative hypothesis when the null hypothesis of no effect is tested.
¶ To test directly whether the paternally and maternally inherited alleles have different effects, their allele frequencies were compared within the cases. Information for this test was mainly captured by the counts of the two types of heterozygote: n12 denotes the number of cases who have inherited allele 1 from the father and allele 2 from the mother, and n21 denotes the number of cases who have inherited allele 2 from the father and allele 1 from the mother.

case-control analysis separately for the paternally and maternally inherited alleles. A 2-d.f. test was applied to evaluate the joint effect. A multiplicative model was assumed for the two alleles, but the magnitude and direction of the effect were allowed to differ. Finally, the difference between the effects of the paternally and maternally inherited alleles was directly tested by comparing their allele frequencies within cases. The information for this test came mainly from the counts of the two types of heterozygote within cases (Supplementary Information).

Two of the seven associations examined, one with prostate cancer and another with coronary artery disease, did not exhibit parental-origin-specific effects (Supplementary Information and Supplementary Table 3). The five associations that did are presented here.

**Breast cancer.** Allele C of rs3817198 in the 11p15 region (Fig. 2) was reported[8] to be associated with breast cancer with an allelic odds ratio of OR = 1.07 ($P = 3 \times 10^{-9}$). This study included about 21,860 cases and 22,578 controls, allowing this modest effect to achieve genome-wide significance. A study[9] of 9,770 cases and 10,799 controls in the Cancer Genetic Markers of Susceptibility project reported odds ratios of 1.02 and 1.12 for heterozygous and homozygous carriers of the same variant, respectively. Using information in their supplementary material, we deduced a *P* value of 0.06. Marker rs3817198 is not on the Illumina chips used to type the majority of the Icelandic samples, but is included on the Illumina 1M BeadChips for which we have data on 124 trios. We used a single-track assay to type another 90 trios, giving a total of 214 trios with genotypes for rs3817198, which translates to a training set of 856 haplotypes. Adapting the statistical model used by IMPUTE[10], allele probabilities of rs3817198 were calculated for individuals with phased and parental-origin-determined haplotypes for this region (Supplementary Information). With the imputation results for 1,803 cases and 34,909 controls (Table 1), the standard case-control test gave a non-significant odds ratio of 1.04 ($P = 0.36$). However, when parental origin was taken into account, the paternally inherited allele showed a significant association (OR = 1.17, $P = 0.0038$). The direct test of parental-origin-specific effects that used only the case data was even more significant ($P = 6.2 \times 10^{-4}$). This is because the estimated effect of allele C when maternally inherited, although not significant ($P = 0.11$), is protective (OR = 0.91).

**Basal-cell carcinoma.** We recently identified association of allele T of rs157935, located at 7q32 (Fig. 3), with basal-cell carcinoma (OR = 1.23, $P = 5.7 \times 10^{-10}$)[7]. Limiting the analysis to samples for which parental origin could be determined, the paternally inherited allele was significantly associated with the disease (OR = 1.40, $P = 1.5 \times 10^{-6}$), but the effect of the maternally inherited allele, although it was in the same direction, was not significant (OR = 1.09, $P = 0.19$; Table 1). Tested directly, the effects of the paternally and maternally inherited alleles were significantly different ($P = 0.01$).

**Type 2 diabetes.** Allele C of rs2237892 in the maternally expressed gene *KCNQ1* was first observed to be associated with T2D in Asian populations[11,12]. The power to detect association in populations of European ancestry is low owing to the high frequency of the variant there (~93% compared with ~61% in Asians), but the association has nonetheless been replicated[11,12]. In the T2D samples we have previously used in genome scans (Table 1) including 1,468 cases, none of the tests involving parental origin were significant for rs2237892. However, with the addition of another 783 patients, giving a total of 2,251 cases (Supplementary Information), allele C was significantly associated with the disease (OR = 1.30, $P = 0.0084$) when maternally transmitted, whereas the results for the paternally inherited allele were flat (OR = 1.03, $P = 0.71$).

Through a meta-analysis of eight T2D genome-wide scans of DIAGRAM sample sets with additional follow-up (Supplementary Information), allele C of rs231362 was shown to associate with the disease (OR = 1.08, $P = 3 \times 10^{-13}$). Marker rs231362 is also located in *KCNQ1* (Fig. 2), but it is not substantially correlated with rs2237892 (correlation coefficient, $r^2 = 0.002$). Also, it is not on any of the Illumina chips used. A training set of 912 haplotypes, created through single-track-assay genotyping of 228 trios, was used for imputation of rs231362 into the Icelandic samples. Using the imputed results, the standard case-control test gave an odds ratio

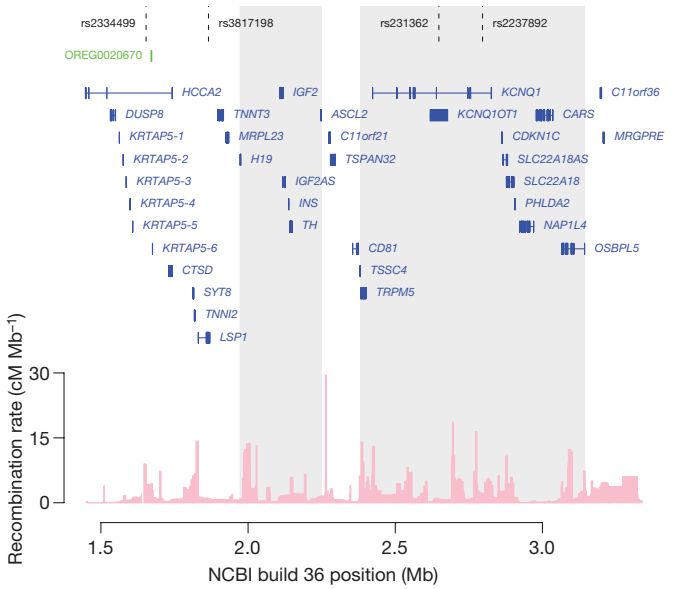



**Figure 2 | Chromosome 11p15 locus.** Markers associated with T2D (rs2334499, rs231362 and rs2237892) and breast cancer (rs3817198) are indicated. The two regions containing clusters of imprinted genes are shaded. The location of the CTCF-binding region studied (OREG0020670) and gene annotations were taken from the University of California, Santa Cruz, genome browser (http://genome.ucsc.edu/). Estimated recombination rates, from the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/), are plotted to reflect the linkage disequilibrium structure in the region. Mb, megabase.

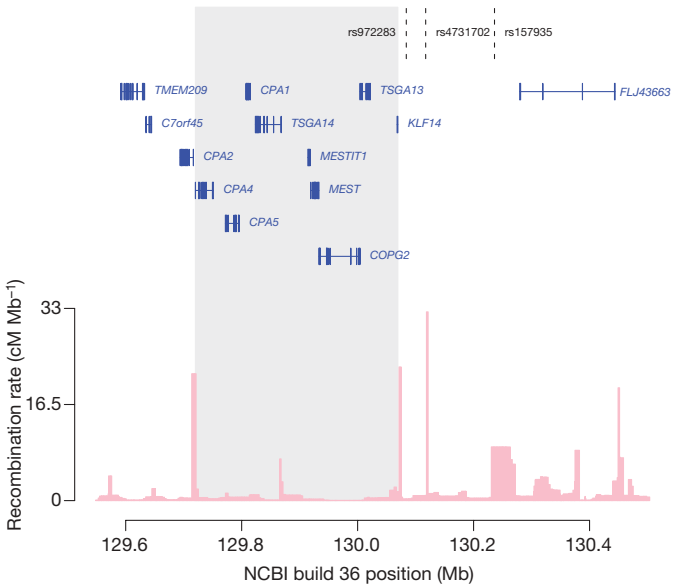



**Figure 3 | Chromosome 7q32 locus.** Markers associated with T2D (rs4731702 and rs972283 ($r^2 = 1$; HapMap CEU)) and basal-cell carcinoma (rs157935) are indicated. The region containing the known imprinted genes is shaded. Gene annotations were taken from the University of California, Santa Cruz, genome browser. Estimated recombination rates (from HapMap) are plotted to reflect the linkage disequilibrium structure in the region.

of 1.10 ($P = 0.013$). The effect, however, appears to be limited to the maternally inherited allele (OR = 1.23, $P = 6.2 \times 10^{-5}$).

Another association with T2D in DIAGRAM samples involves allele C of rs4731702 at 7q32 (OR = 1.07, $P = 2 \times 10^{-10}$; Fig. 3). In our combined Icelandic samples, the association was again restricted to the maternally inherited allele (OR = 1.17, $P = 0.0010$; OR = 0.99, $P = 0.79$ for the paternally inherited allele).

Evaluating the seven known susceptibility variants jointly (the five highlighted above plus the two variants for prostate cancer and coronary artery disease), the test for no parental-specific effect for all gave a $P$ value of $<5 \times 10^{-6}$. Also, an analysis of false-discovery rate[13] indicates that it is likely that at least four of the five highlighted variants have true parental-origin-specific effects (Supplementary Information).

**A new diabetes susceptibility variant.** Properly evaluating the statistical significance of the susceptibility variants described above requires adjusting for relatedness of the participants using the method of genomic control[14]. This required us to perform genome scans for these diseases (Supplementary Table 4 gives parental-origin test results for established susceptibility variants located outside the selected regions). The T2D scan performed with the initial sample set (Supplementary Information and Supplementary Fig. 2) gave a striking result (Table 1). Allele T of rs2334499, at 11p15 (Fig. 2), showed such a weak association (OR = 1.11, $P = 0.017$) in the standard case-control test that it does not stand out in a genome-wide scan. However, taking into account parental origin, both the paternally inherited allele (OR = 1.41, $P = 4.3 \times 10^{-9}$) and the 2-d.f. test ($P = 3.5 \times 10^{-9}$) were genome-wide significant. Most notably, the maternally inherited allele also showed nominally significant association, but the effect of allele T was protective (OR = 0.87, $P = 0.020$). Tested directly, the difference between the effects of the paternally and maternally inherited alleles was also genome-wide significant ($P = 7.0 \times 10^{-9}$). This SNP falls within 350 kb of a large cluster of imprinted genes, making the results even more compelling. However, the observation that allele T is protective when maternally inherited required replication. For this, we used an additional set of 783 chip-typed T2D cases. All tests involving parental origin were significantly replicated. For the combined analysis of the two sample sets (Supplementary Information and Supplementary Fig. 3), the paternally inherited allele had an odds ratio of 1.35 ($P = 4.7 \times 10^{-10}$) and the maternally inherited allele had an odds ratio of 0.86 ($P = 0.0020$). The 2-d.f. test and the paternal-versus-maternal test gave $P$ values of $5.7 \times 10^{-11}$ and $4.1 \times 10^{-11}$, respectively.

As there are known examples in an imprinted setting where the paternal and maternal alleles interact[15], we tested rs2334499 for an interactive effect. This test was not significant ($P > 0.4$; Supplementary Information) indicating that the multiplicative model provides an adequate fit. Specifically, in comparison with CT (first allele paternal, second allele maternal), CC, TT and TC have relative risks of 1.17, 1.35 and 1.57, respectively.

The transmitted maternal allele has an effect in all four T2D variants in Table 1. Because prenatal maternal conditions may be a factor in conferring risk on the offspring, we examined the role of the non-transmitted maternal allele. No significant effect was observed (Supplementary Information).

## Imprinted regions at 11p15 and 7q32

Imprinted genes at 11p15.5 fall into two clusters, *H19*/*IGF2* and *KCNQ1* (Fig. 2), regulated through separate imprinting control regions, each of which controls expression of a number of genes within the cluster[16]. The *H19*/*IGF2* imprinting control region is regulated through a differentially methylated region that is normally methylated only on the paternal chromosome. Binding of the insulator protein CTCF in the imprinting control region is permitted only on the unmethylated maternal chromosome, resulting in expression of *IGF2* only from the paternal methylated chromosome and expression of *H19* from the maternal chromosome[17]. The breast cancer paternally associated marker rs3817198 resides within *LSP1*, 100 kb downstream of *H19* and within the same linkage disequilibrium block. The effect of this marker on breast cancer could thus be through the *H19*/*IGF2* imprinted locus. Loss of imprinting at the *H19*/*IGF2* locus, resulting in activation of *IGF2* expression, has been reported in a number of different tumour types[18]. Furthermore, loss of imprinting at the *H19*/*IGF2* locus in normal tissue has also been shown to indicate a predisposition to colorectal cancer[18].

The *KCNQ1* cluster is regulated through an imprinting control region located in the promoter region of *KCNQ1OT1*, a paternally expressed non-coding antisense RNA. Hypermethylation of the maternal allele results in monoallelic activity of the neighbouring maternally expressed protein-coding genes. The two T2D-associated markers at this locus, rs231362 and rs2237892, are both located within the maternally expressed *KCNQ1*, consistent with the risk associations, rs231362 also residing within the *KCNQ1OT1* antisense transcript (Fig. 2).

Although both the T2D marker rs2334499 and the breast cancer marker rs3817198 fall within 350 kb of imprinted genes, the region harbouring them has not been reported to be imprinted[19] (Fig. 2). Marker rs2334499 resides within the first intron of *HCCA2*, a gene which spans 300 kb containing several other genes (Fig. 2) including *KRTAP5-1* to *KRTAP5-6*, *DUSP8* and *CTSD*[20]. To determine whether genes in this region showed signs of imprinting, we performed allele-specific expression analysis of *HCCA2*, *CTSD* and *DUSP8* (Fig. 2), as well as three genes known to be imprinted in the 11p15.5 region (*IGF2*, *KCNQ1* and *KCNQ1OT1*), in RNA isolated from peripheral blood and adipose. Whereas allele-specific expression of *IGF2*, *KCNQ1* and *KCNQ1OT1* was confirmed in this data set, clear biallelic expression was seen for *HCCA2* and *DUSP8*. However, excess paternal expression could not be ruled out for *CTSD* (Supplementary Information and Supplementary Table 6).

The imprinted region at 7q32 consists of maternally expressed genes (*CPA4* and *KLF14*) flanking paternally expressed genes (*MEST* and *MESTIT1*) (Fig. 3). The T2D-associated marker rs4731702 is located 14 kb from the maternally expressed *KLF14* transcription factor[21] and only increases risk of T2D when carried on the maternal chromosome. The basal-cell carcinoma variant rs157935, conferring risk through the paternal allele, is located 170 kb telomeric to the imprinted region.

We previously[22] correlated SNP genotypes from the Illumina 300K chip with gene expression using RNA samples from adipose tissue ($N = 603$) and peripheral blood ($N = 745$). Here, taking parental origin into account, we re-evaluated the correlation between the six variants in Table 1 and expression of genes at the 7q32 and 11p15.5 loci. The T2D risk allele of rs4731702 at 7q32 correlated with lower expression of *KLF14* in adipose tissue ($P = 3 \times 10^{-21}$) when inherited maternally, but there was no effect when it was inherited paternally (Supplementary Table 7). Similar correlation was not seen in blood. Conversely, no strong correlation with parental-origin-specific gene expression was seen for the other disease-associated variants at 7q32 or 11p15.5 (Supplementary Table 7).

## Methylation of a novel CTCF-binding site

Recent studies have mapped regions of CTCF-binding genome-wide for identification of insulator elements[23,24]. One of the sites identified (OREG0020670) is a 2-kb region located 17 kb centromeric to the T2D marker rs2334499 (Fig. 2 and Supplementary Fig. 4). We assessed the methylation status of this CTCF-binding region in DNA samples derived from peripheral blood, using bisulphite sequencing. We identified a differentially methylated region of 180 base pairs including seven CpG dinucleotides (Supplementary Fig. 4) where the ratio of 5-methyl cytosine (Cp) varied from around 0.1 to 0.6. Methylation at five of the seven CpG dinucleotides (CpG-1 to CpG-5; Supplementary Fig. 4) was highly correlated (Supplementary Table 9). The estimated Cp ratio was tested for correlation with SNPs in a two-megabase surrounding region. The most significant correlation was observed between methylation status at CpG-4 and

**Table 2 | Correlation between methylation of a CTCF-binding region and the T2D risk-variant rs2334499**

| CpG dinucleotide | Percentage methylation (mean, s.e.)* | Effect† | $P$‡ |
|---|---|---|---|
| CpG-1 | 22.9 (0.9) | −5.7 | $6.5 \times 10^{-7}$ |
| CpG-2 | 15.3 (0.7) | −3.1 | 0.00055 |
| CpG-3 | 13.3 (0.8) | −2.5 | 0.017 |
| CpG-4 | 56.7 (0.9) | −8.4 | $2.6 \times 10^{-13}$ |
| CpG-5 | 34.9 (0.9) | −6.7 | $6.8 \times 10^{-8}$ |
| CpG-6 | 22.2 (0.6) | −1.0 | 0.24 |
| CpG-7 | 52.9 (1.1) | −1.5 | 0.30 |

* Mean and standard error of the C/T allele ratio estimated by bisulphite sequencing of 168 individuals.
† Change in percentage methylation per allele T of rs2334499 carried.
‡ Significance of the correlation between methylation and rs2334499.

rs2334499, for which $P = 2.6 \times 10^{-13}$ (Table 2). Furthermore, correlation between rs2334499 and methylation of CpG-1 to CpG-5 was significant. For these five CpG dinucleotides, the T2D risk allele correlated with decreased methylation and this effect was observed regardless of whether the allele was inherited from the father or the mother. By contrast, neither the breast cancer variant nor the two other T2D markers at 11p15.5 showed any correlation with the methylation status of this CTCF-binding site.

## Discussion

Being able to determine parental origin of alleles and haplotypes in large samples opens new avenues to study associations between sequence variants and human traits. Standard association analysis provides suboptimal power to discover disease susceptibility variants that exhibit parental-origin-specific effects. Even when association can be established, the true effect is underestimated. Marker rs2334499 did not gain serious attention even after the large collaborative effort of the DIAGRAM Consortium. However, its contribution to T2D, measured by the recurrent risks of siblings generated, is second only to that of the *TCF7L2* variant among the known susceptibility variants (Supplementary Information and Supplementary Fig. 2). Sequence variants, such as rs2334499, that can confer both risk and protection depending on parental origin can lead to balanced selection and as a result promote diversity.

Functional imprinting is extremely tissue and stage specific, and although some genes retain their imprinted status throughout life, the main role of imprinting is believed to be during prenatal growth and development. However, the associations of rs4731702 C with T2D and *KLF14* expression in adult adipose tissue, in both cases only when maternally inherited, strongly implicates this transcription factor as the disease gene.

We searched for evidence of epigenetic marks around the T2D risk variant rs2334499, as it is located some distance away from the established 11p15.5 imprinted genes. A CTCF-binding site in the region was found to be differentially methylated and the rs2334499 risk allele was shown to be correlated with decreased methylation. Given the well-established role of CTCF in imprinting, this new site could differentially influence the dosage of imprinted genes on the two parental chromosomes.

Despite their successes, genome-wide association studies have so far yielded sequence variants that explain only a small fraction of the estimated heritability of most of the human traits studied. Obvious contributors to the unexplained heritability, or 'dark matter', include rare variants not well tagged by common SNPs and common variants that have very small effects individually. Results presented here demonstrate that a portion of the heritability of some common/complex traits is hidden in more complex relations between sequence variants and the risks of these variants.

## METHODS SUMMARY

**Subjects.** We used 38,167 Icelandic individuals who were genotyped using an Illumina SNP chip and processed for long-range phasing. See Supplementary Information for details of disease and control groups.

**Genotyping.** We performed genome-wide genotyping using various Illumina BeadChips. Individual genotyping of two SNPs was done using Centaurus assays.

**Determination of parental origin.** We used the Icelandic genealogy database to identify the closest relatives who shared a haplotype with the proband. Parental origin was then assigned to the two haplotypes of a proband on the basis of a computed score (Methods).

**Data imputation.** On the basis of a training set of trios, by adapting the statistical model used by IMPUTE[10] to our setting, we computed allele probabilities of the paternal and maternal chromosomes separately for samples for which an SNP was not genotyped.

**Statistical analysis.** For SNPs directly typed, we used likelihood-based procedures to study disease associations taking parental origin into account. For imputed SNPs, we used logistic regressions and *t*-tests. Genomic control was used to control for relatedness among subjects.

**Methylation analysis.** Bisulphite sequencing was used to estimate the level of methylation. For each CpG dinucleotide, we determined the methylation status by calculating the C/T allele ratio at that site.

**Gene expression.** We tested associated SNPs for correlation with expression of genes located in a one-megabase window centred on the variant, in a data set of expression measurements in whole blood and adipose tissue. The same expression library was used to determine parental origin of expression by using allele-specific probes where the parental origin of each allele was known.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 14 August; accepted 29 October 2009.**


1. Rampersaud, E., Mitchell, B. D., Naj, A. C. & Pollin, T. I. Investigating parent of origin effects in studies of type 2 diabetes and obesity. *Curr. Diabetes Rev.* **4,** 329–339 (2008).
2. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40,** 1068–1075 (2008).
3. Luedi, P. P. *et al.* Computational and experimental identification of novel human imprinted genes. *Genome Res.* **17,** 1723–1730 (2007).
4. Morison, I. M., Paton, C. J. & Cleverley, S. D. The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res.* **29,** 275–276 (2001).
5. Morison, I. M., Ramsay, J. P. & Spencer, H. G. A census of mammalian imprinting. *Trends Genet.* **21,** 457–465 (2005).
6. Hindorff, L. A., Junkins, H. A., Mehta, J. P. & Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. *OPG: Catalog Published Genome-Wide Assoc. Studies* ⟨http://www.genome.gov/gwastudies⟩ (2009).
7. Stacey, S. N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nature Genet.* **41,** 909–914 (2009).
8. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).
9. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*). *Nature Genet.* **41,** 579–584 (2009).
10. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39,** 906–913 (2007).
11. Yasuda, K. *et al.* Variants in *KCNQ1* are associated with susceptibility to type 2 diabetes mellitus. *Nature Genet.* **40,** 1092–1097 (2008).
12. Unoki, H. *et al.* SNPs in *KCNQ1* are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature Genet.* **40,** 1098–1102 (2008).
13. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
14. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
15. Georges, M., Charlier, C. & Cockett, N. The callipyge locus: evidence for the *trans* interaction of reciprocally imprinted genes. *Trends Genet.* **19,** 248–252 (2003).
16. Ideraabdullah, F. Y., Vigneau, S. & Bartolomei, M. S. Genomic imprinting mechanisms in mammals. *Mutat. Res.* **647,** 77–85 (2008).
17. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405,** 482–485 (2000).
18. Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447,** 433–440 (2007).
19. Goldberg, M., Wei, M., Yuan, L., Murty, V. V. & Tycko, B. Biallelic expression of HRAS and MUCDHL in human and mouse. *Hum. Genet.* **112,** 334–342 (2003).
20. Authier, F., Metioui, M., Fabrega, S., Kouach, M. & Briand, G. Endosomal proteolysis of internalized insulin at the C-terminal region of the B chain by cathepsin D. *J. Biol. Chem.* **277,** 9437–9446 (2002).
21. Parker-Katiraee, L. *et al.* Identification of the imprinted *KLF14* transcription factor undergoing human-specific accelerated evolution. *PLoS Genet.* **3,** e65 (2007).
22. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
23. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128,** 1231–1245 (2007).

24. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* **19,** 24–32 (2009).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** This project was funded in part by FP7-MC-IAPP Grant agreement no. 218071 (CancerGene) to deCODE genetics.

**Author Contributions** A.K. and K.S. planned and directed the research. A.K. wrote the first draft of the paper and, together with K.S., V.S., G.M., G.T. and U.T., wrote most of the final version. A.K. and G.M. designed the method to determine parental origin. G.M., with assistance from P.I.O., implemented the algorithm. D.F.G. wrote the code for association analysis taking parental origin into account and performed some initial analyses. P.S., S.B. and S.S. tabulated the established disease-associated variants and the regions known to harbour imprinted genes. V.S. and G.T. contributed to the analysis of the diabetes data and, together with A.K. and U.T., planned the follow-up association and functional studies. A.G., A.K. and M.L.F. imputed the untyped SNPs. S.N.S. and P.S. were responsible for the breast cancer and basal-cell carcinoma data. A.B.H., G.S. and R.B. provided clinical data for T2D, O.Th.J., T.J. and H.S. provided clinical data for breast cancer, and J.H.O., B.S. and K.R.B. provided clinical data for basal-cell carcinoma. The DIAGRAM Consortium provided the novel T2D-associated variants that are close to imprinted genes. Aslaug J., A.S., Adalbjorg J., K.Th.K. and S.A.G. performed the methylation and expression studies. A.C.F.-S. assisted in the interpretation of the results from the association and functional studies.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.K. (kong@decode.is) or K.S. (kstefans@decode.is).

**DIAGRAM Consortium** Benjamin F. Voight[1,2,3], Laura J. Scott[4], Valgerdur Steinthorsdottir[5], Christian Dina[6,7], Eleftheria Zeggini[8,9], Cornelia Huth[10,11], Yurii S. Aulchenko[12], Ryan P. Welch[4], Gudmar Thorleifsson[5], Laura J. McCulloch[13], Teresa Ferreira[9], Harald Grallert[10,11], Najaf Amin[12], Guanming Wu[14], Cristen J. Willer[4], Soumya Raychaudhuri[1,2,15], Shaun Purcell[1,2,16], Steve A. McCarroll[1,17], Claudia Langenberg[18], Oliver M. Hoffmann[19], Josée Dupuis[20], Lu Qi[21,22], Ayellet V. Segrè[1,17], Mandy van Hoek[23], Pau Navarro[24], Kristin Ardlie[1], Beverley Balkau[25,26], Rafn Benediktsson[27,28], Amanda J. Bennett[13], Roza Blagieva[29], Eric Boerwinkle[30], Lori L. Bonnycastle[31], Kristina Bengtsson Boström[33], Bert Bravenboer[34], Suzannah Bumpstead[8], Noël P. Burtt[1], Guillaume Charpentier[35], Peter S. Chines[31], Marilyn Cornelis[22], David J. Couper[36], Gabe Crawford[1], Alex S. F. Doney[37,38], Katherine S. Elliott[9], Amanda L. Elliott[1,17], Michael R. Erdos[31], Caroline S. Fox[39,40], Christopher S. Franklin[41], Martha Ganser[4], Christian Gieger[10], Niels Grarup[42], Todd Green[1,2], Simon Griffin[18], Christopher J. Groves[13], Candace Guiducci[1], Samy Hadjadj[43], Neelam Hassanali[13], Christian Herder[44], Bo Isomaa[45,46], Anne U. Jackson[4], Paul R. V. Johnson[47], Torben Jørgensen[48], Wen H. L. Kao[49], Norman Klopp[10], Augustine Kong[5], Peter Kraft[21], Johanna Kuusisto[50], Torsten Lauritzen[51], Man Li[52], Alouisius Lieverse[53], Cecilia M. Lindgren[9], Valeriya Lyssenko[54], Michel Marre[55,56], Thomas Meitinger[10], Kristian Midthjell[57], Mario A Morken[31], Narisu Narisu[31], Peter Nilsson[54], Katharine R. Owen[13], Felicity Payne[8], John R. B. Perry[58,59], Ann-Kristin Petersen[10], Carl Platou[57], Christine Proença[6], Inga Prokopenko[9,13], Wolfgang Rathmann[60], N. William Rayner[9,13], Neil R. Robertson[9,13], Ghislain Rocheleau[61,62,63], Michael Roden[44,64], Michael J. Sampson[65], Richa Saxena[1,2,66], Beverley M. Shields[58,59], Peter Shrader[67,68], Gunnar Sigurdsson[27,28], Nicholas Smith[8], Thomas Sparsø[42], Klaus Strassburger[60], Heather M. Stringham[4], Qi Sun[21], Amy J. Swift[31], Barbara Thorand[10], Jean Tichet[69], Tiinamaija Tuomi[45,70], Rob van Dam[22], Thijs van Herpt[23,53], G. Bragi Walters[5], Michael N. Weedon[58,59], Jacqueline Witteman[12], Richard N. Bergman[71], Stephane Cauchi[6], Francis S. Collins[72], Anna L. Gloyn[13], Ulf Gyllensten[73], Torben Hansen[42,74], Winston A. Hide[19], Graham A. Hitman[75], Albert Hofman[12], David Hunter[21], Kristian Hveem[57,76], Markku Laakso[50], Karen L. Mohlke[77], Andrew D. Morris[37,38], Colin N. A. Palmer[37,38], Peter P. Pramstaller[78], Igor Rudan[41,79,80], Eric Sijbrands[23], Lincoln D. Stein[14], Jaakko Tuomilehto[81], Andre Uitterlinden[23], Mark Walker[82], Nicholas J. Wareham[18], Richard M. Watanabe[83], Goncalo R. Abecasis[4], Inês Barroso[8], Bernhard O. Boehm[29], Harry Campbell[41], Mark J. Daly[1,2], Jose C. Florez[1,2,3], Timothy M. Frayling[58,59], Leif Groop[54,70], Andrew T. Hattersley[58,59], Frank B. Hu[21,22], James B. Meigs[3,67], Andrew P. Morris[9], James S. Pankow[84], Oluf Pedersen[42,85,86], Rob Sladek[61,62,63], Unnur Thorsteinsdottir[5,87], H.-Erich Wichmann[10,11], James F. Wilson[41], Thomas Illig[10], Philippe Froguel[6,88], Cornelia M. van Duijn[12], Kari Stefansson[5,87], David Altshuler[1,2,3,17,66,89], Michael Boehnke[4], Mark I. McCarthy[9,13,90].

[1]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [3]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [4]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA. [5]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland. [6]CNRS-UMR-8090, Institute of Biology and Lille 2 University, Pasteur Institute, F-59019 Lille, France. [7]INSERM, UMR915, CNRS, ERL3147, 44007 Nantes, France. [8]Wellcome Trust Sanger Institute, Hinxton CB10 1HH, UK. [9]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [10]Institute of Epidemiology, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany. [11]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany. [12]Department of Epidemiology, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [13]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LJ, UK. [14]Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario M5G 0A3, Canada. [15]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [16]Stanley Center for Psychiatric Research, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [17]Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [18]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [19]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [20]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02118, USA. [21]Departments of Nutrition and Epidemiology, Harvard School of Public Health, 665 Huntington Avenue, Boston, Massachusetts 02115, USA. [22]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, Massachusetts 02115, USA. [23]Department of Internal Medicine, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands. [24]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, UK. [25]INSERM U780, F-94807 Villejuif, France. [26]University Paris-Sud, F-91405 Orsay, France. [27]Landspitali University Hospital, 101 Reykjavik, Iceland. [28]Icelandic Heart Association, 201 Kopavogur, Iceland. [29]Division of Endocrinology, Diabetes and Metabolism, Ulm University, 89081 Ulm, Germany. [30]The Human Genetics Center and Institute of Molecular Medicine, University of Texas Health Science Center, Houston, Texas 77030, USA. [31]National Human Genome Research Institute, National Institute of Health, Bethesda, Maryland 20892, USA. [33]R&D Centre, Skaraborg Institute, 541 30 Skövde, Sweden. [34]Department of Internal Medicine, Catharina Hospital, PO Box 1350, 5602 ZA Eindhoven, The Netherlands. [35]Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, F-91100 Corbeil-Essonnes, France. [36]Department of Biostatistics and Collaborative Studies Coordinating Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [37]Diabetes Research Centre, [38]Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee DD1 9SY, UK. [39]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts 01702, USA. [40]Division of Endocrinology, Diabetes, and Hypertension, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [41]Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. [42]Hagedorn Research Institute, DK-2820 Gentofte, Denmark. [43]CHU de Poitiers, Endocrinologie Diabetologie, CIC INSERM 0801, INSERM U927, Université de Poitiers, UFR, Médecine Pharmacie, 86021 Poitiers Cedex, France. [44]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [45]Folkhälsan Research Center, FIN-00014 Helsinki, Finland. [46]Malmska Municipal Health Center and Hospital, 68601 Jakobstad, Finland. [47]DRWF Human Islet Isolation Facility and Oxford Islet Transplant Programme, University of Oxford, Old Road, Headington, Oxford OX3 7LJ, UK. [48]Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark. [49]Department of Epidemiology, Department of Medicine, and Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University, Baltimore, Maryland 21287, USA. [50]Department of Medicine, University of Kuopio and Kuopio University Hospital, FIN-70211 Kuopio, Finland. [51]Department of General Medical Practice, University of Aarhus, DK-8000 Aarhus, Denmark. [52]Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21287, USA. [53]Department of Internal Medicine, Maxima MC, PO-Box 90052, 5600 PD Eindhoven, The Netherlands. [54]Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, 205 02 Malmö, Sweden. [55]Department of Endocrinology, Diabetology, Nutrition, Bichat-Claude Bernard University Hospital, Assistance Publique des Hôpitaux de Paris, 75877 Paris Cedex 18, France. [56]INSERM U695, Université Paris 7, 75870 Paris Cedex 18, France. [57]HUNT Research Center, Department of Community Medicine and General Practice, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway. [58]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK. [59]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Barrack Road, Exeter EX2 5DW, UK. [60]Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [61]Department of Human Genetics, McGill University, Montreal H3H 1P3, Canada. [62]Department of Medicine, Faculty of Medicine, McGill University, Montreal H3A 1A4, Canada. [63]McGill University and Genome Quebec Innovation Centre, Montreal H3A 1A4, Canada. [64]Department of Medicine/Metabolic Diseases, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [65]Department of Endocrinology and Diabetes, Norfolk and Norwich University Hospital NHS Trust , Norwich NR1 7UY, UK. [66]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [67]General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts, USA. [68]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. [69]Institut Interrégional pour la Santé, F-37521 La Riche, France. [70]Department of Medicine, Helsinki University Hospital, University of Helsinki, FIN-00290 Helsinki, Finland. [71]Department of Physiology and Biophysics, University of Southern California School of Medicine, Los Angeles, California

90033, USA. [72]National Institutes of Health, Bethesda, Maryland 20892, USA. [73]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, S-751 85 Uppsala, Sweden. [74]University of Southern Denmark, DK-5230 Odense, Denmark. [75]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [76]Department of Medicine, The Hospital of Levanger, N-7600 Levanger, Norway. [77]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA. [78]Institute of Genetic Medicine, European Academy Bozen/Bolzano, Viale Druso 1, 39100 Bolzano, Italy. [79]Croatian Centre for Global Health, Faculty of Medicine, University of Split, Soltanska 2, 21000 Split, Croatia. [80]Institute for Clinical Medical Research, University Hospital ''Sestre Milosrdnice'', Vinogradska 29, 10000 Zagreb, Croatia. [81]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki FIN-00300, Finland. [82]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [83]Department of Preventive Medicine, Keck Medical School, University of Southern California, Los Angeles, California 90089-9001, USA. [84]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota 55454, USA. [85]Department of Biomedical Science, Panum, Faculty of Health Science, University of Copenhagen, 2200 Copenhagen, Denmark. [86]Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark. [87]Faculty of Medicine, University of Iceland, 101 Reykjavík, Iceland. [88]Genomic Medicine, Imperial College London, Hammersmith Hospital, London W12 0NN, UK. [89]Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02144, USA. [90]Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK.

## METHODS

**Assignment of parental origin.** Let $H$ be a haplotype for a tile $T$. For a particular proband, $f(T, H)$ and $m(T, H)$ were calculated as the meiotic distances to the closest relatives on the paternal side and, respectively, the maternal side known to carry $H$. Descendants of the parents of the proband, for example siblings of the proband, were excluded from this calculation. Also, a value of 10,000 was assigned when no relatives carrying the haplotype was found. Let $A$ and $B$ be the two phased haplotypes of the proband. The single-tile score for parental origin was calculated as

$$\begin{aligned}\text{score}(T,A,B) &= \text{score}(T,A) - \text{score}(T,B)\\ &= [\log(1-2^{-m(T,A)}) - \log(1-2^{-f(T,A)})]\\ &\quad - [\log(1-2^{-m(T,B)}) - \log(1-2^{-f(T,B)})]\end{aligned}$$

A score that is greater than zero supports the assignment of $A$ as the paternally inherited haplotype and $B$ as the maternally inherited haplotype, whereas a score that is less than zero supports the reverse. Although it is not meant to be optimal in any formal sense, this system of scoring was chosen to have two properties. First, for the same absolute difference between $m(T, H)$ and $f(T, H)$, the absolute value of score$(T, H)$ is higher when the lesser of $m(T, H)$ and $f(T, H)$ is smaller, thus giving more weight to situations in which a close relative who shared a haplotype is found. Second, the scoring was designed such that the result from one haplotype in one tile could not completely dominate the contributions from other haplotypes and adjacent tiles when results were combined (see below).

When haplotypes for $n$ consecutive tiles, $T_1, \ldots, T_n$, could be stitched together to form $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_n)$, the contig score for parental origin assignment was calculated as

$$\text{contig\_score}(T_1, \ldots, T_n) = \sum_{i=1}^{n} \text{score}(T_i)$$

Parental origins were assigned on the basis of whether the contig score was greater than or less than zero. The accuracy of this procedure was evaluated using the trio test.

ARTICLES

# Growth landscape formed by perception and import of glucose in yeast

Hyun Youk[1] & Alexander van Oudenaarden[1,2]

**An important challenge in systems biology is to quantitatively describe microbial growth using a few measurable parameters that capture the essence of this complex phenomenon. Two key events at the cell membrane—extracellular glucose sensing and uptake—initiate the budding yeast's growth on glucose. However, conventional growth models focus almost exclusively on glucose uptake. Here we present results from growth-rate experiments that cannot be explained by focusing on glucose uptake alone. By imposing a glucose uptake rate independent of the sensed extracellular glucose level, we show that despite increasing both the sensed glucose concentration and uptake rate, the cell's growth rate can decrease or even approach zero. We resolve this puzzle by showing that the interaction between glucose perception and import, not their individual actions, determines the central features of growth, and characterize this interaction using a quantitative model. Disrupting this interaction by knocking out two key glucose sensors significantly changes the cell's growth rate, yet uptake rates are unchanged. This is due to a decrease in burden that glucose perception places on the cells. Our work shows that glucose perception and import are separate and pivotal modules of yeast growth, the interaction of which can be precisely tuned and measured.**

In 1942 Jacques Monod introduced his microbial growth model[1] that prompted quantitative studies of microbial metabolism[2–13]. This motivated a wealth of mathematical models describing the growth of budding yeast *Saccharomyces cerevisiae* on the key carbohydrate glucose[14]. These models mainly focus on the effect of glucose import on the growth rate. However, in addition to importing glucose, yeast senses extracellular glucose through several glucose sensors. These two key events at the cell membrane—glucose sensing and import—then trigger many downstream intracellular molecular events (for example, transcription, metabolic processes, post-transcriptional modifications) that collectively determine the growth rate[15]. Many conventional models overlook this collective effect by ignoring glucose sensing. Growth behaviours that are qualitatively very different from current models' descriptions may arise if glucose sensing and import are properly taken into account. One approach to addressing this deficiency is constructing detailed many-parameter models that attempt to explicitly track each of the vast molecular events involved in yeast's glucose metabolism[8,9]. Such an approach has provided detailed information on the flux of thousands of known metabolic reactions and new insights into yeast's growth on glucose. However, it also combines the effects of glucose sensing and import because it is not yet known how each of the vast molecular events are altered when glucose import rate is varied independently of the level of extracellular glucose sensed by the cell. The enormous number of metabolites and reactions involved makes experimentally determining each molecular change due to glucose sensing and import challenging. Indeed, a persistent challenge in obtaining a quantitative understanding of microbial growth on nutrients has been identifying just the few parameters that are necessary for extracting the central features from this complex cellular process. A phenomenological model that retains just those essential parameters may provide new insights and central design principles[16,17] underlying microbial growth. Motivated by these considerations, we sought to decouple and measure the separate effects of glucose sensing and import on cell growth, then provide a concise phenomenological model that determines how the interaction between the two determines the growth rate.

## Dependence of growth rate on glucose level

To measure and separate out the effects of glucose perception and import on growth rate, we first decouple any control that glucose sensing has on glucose import. Such coupling primarily comes from the two glucose sensors (Snf3 and Rgt2)[18] that drive the transcriptional regulation of the six primary hexose transporters (Hxt1–4, 6 and 7)[19–23] that are responsible for glucose import (Supplementary Fig. 1). Our background strain lacks all the major and minor glucose transporter genes (*hxt1-17*Δ, *agt1*Δ, *stl1*Δ, *gal2*Δ)[24], thus no sensors affect the transcription level of any transporter genes including the *HXT* genes. We made five 'single-*HXT*' strains by introducing one of the five primary *HXT* genes (excluding *HXT7*) into the background strain, under the control of the inducible promoter $P_{TET07}$. Each strain contains just one type of *HXT* gene, and its expression level could be controlled by the inducer doxycycline independently of extracellular glucose (Supplementary Fig. 2).

We measured the log-phase growth rate of the single-*HXT* strains in minimal media containing a range of different concentrations of doxycycline and glucose, the concentrations of which were held constant during batch growth for each experiment. We found surprising behaviours in the growth rate of each single-*HXT* strain (Fig. 1 and Supplementary Fig. 3). Because glucose no longer regulates the transcription of the sole *HXT* gene in our strains in a complicated manner, one would expect that an increase in extracellular glucose concentration would lead to a simple increase in the single-*HXT* strain's glucose uptake rate (when the doxycycline concentration is held constant). A typical conventional model[14] predicts that the growth rate should thus simply rise as the glucose level increases. Yet, depending on the initial glucose level, a further increase either increases or decreases the 'Hxt1-only' strain's growth rate (Fig. 1). This is also true for the growth rates of the Hxt2-only and Hxt4-only

[1]Department of Physics, [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

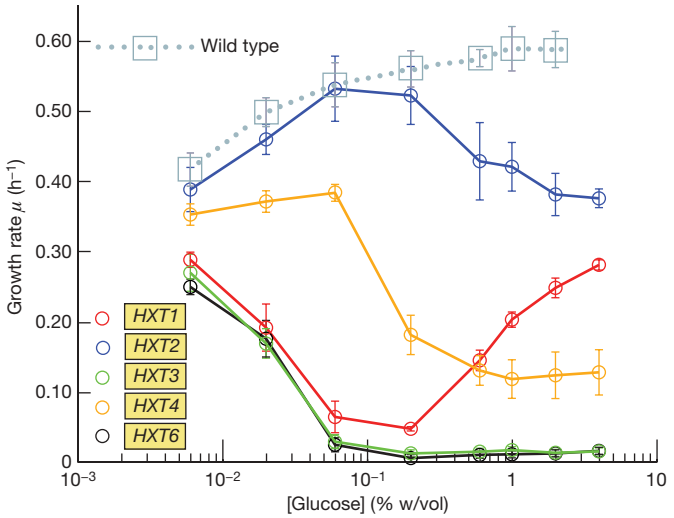



**Figure 1 | Growth rates of single-*HXT* strains do not show any systematic trend with respect to glucose concentration.** Log-phase growth rates of the wild-type strain (CEN.PK2-1C) and five single-*HXT* strains at varying [glucose] but constant [doxycycline] (0 μg ml$^{-1}$ for wild-type and 2.5 μg ml$^{-1}$ for single-*HXT* strains) are shown. The shape of each single-*HXT* strain's growth-rate curve is maintained over a wide range of doxycycline concentrations (Supplementary Fig. 3). The growth-rate curves of the single-*HXT* strains show stark differences from the wild-type's curve: single-*HXT* strains' growth rates can substantially decrease, and some strains even approach growth arrest, despite a monotonic increase in [glucose]. Error bars, s.e.m.; $n = 3$.

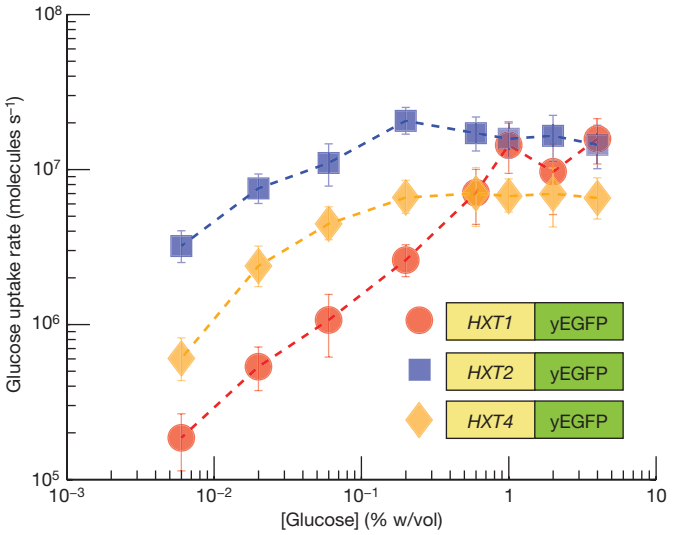



**Figure 2 | A rise in [glucose] produces an increase in the uptake rate, but cells do not necessarily grow faster.** To both measure and calculate glucose uptake rates, yEGFP was fused to the *HXT* gene in each single-*HXT* strain. These fluorescent single-*HXT* strains have the same growth-rate features as their non-fluorescent counterparts shown in Fig. 1 (Supplementary Fig. 4). The measured glucose uptake rates per cell for just three of these fluorescent single-*HXT* strains at [doxycycline] = 2.5 μg ml$^{-1}$ are shown here. These fluorescent single-*HXT* strains' glucose uptake rates monotonically increase as [glucose] increases, despite the non-systematic behaviour of their growth rates reflected in Fig. 1. Hence, a cell can grow faster, or slower, or approach growth arrest despite an increase in both its glucose uptake rate and [glucose]. Error bars, s.e.m.; $n = 3$.

strains. Furthermore, despite growing as well as other strains at low glucose levels, the *HXT3*-only and *HXT6*-only strains even approach growth arrest for glucose level higher than 0.02% (Fig. 1). Thus, we observed no systematic relationship between glucose level and growth rate. It is noteworthy that the wild-type strain, unlike these single-*HXT*s, simply grows faster when more glucose is present (Fig. 1), a behaviour we will consider more closely later.

## Dependence of growth rate on glucose uptake rate

Using our doxycycline-inducible expression system, we were able to show that for every single-*HXT* strain at fixed doxycycline level, the glucose uptake rate increased as the glucose level increased (Fig. 2). To measure glucose uptake rates, we fused yeast-enhanced green fluorescent protein (yEGFP) to the inducible *HXT* gene in each of the single-*HXT* strains (Supplementary Fig. 4). Measuring the average single-cell fluorescence in these strains gave us the relative number of Hxt proteins synthesized in these cells (Supplementary Fig. 5). Using the known Michaelis–Menten parameters of the Hxts[25,26], we calculated the cell's total glucose uptake rate. We also directly measured the cell's glucose uptake rate. The directly measured and calculated uptake rates were in good agreement (Supplementary Fig. 6): glucose uptake rate increased as the glucose concentration increased (at constant doxycycline concentration) (Fig. 2 and Supplementary Fig. 7). Hence, despite a monotonic increase in both glucose uptake rate and extracellular glucose level, single-*HXT* strains at fixed doxycycline concentration can grow significantly faster, or slower, or even approach growth arrest as seen earlier (Fig. 1)—effects that no conventional growth model can either quantitatively or qualitatively describe.

## Phenomenological model of growth

Plotting all five single-*HXT* strains' growth rates and uptake rates together resulted in a wide scatter of data points, in which each data point is specified by two coordinates: uptake rate and growth rate (Fig. 3a and Supplementary Fig. 8). This plot shows that uptake rate alone cannot specify the cell's growth rate. Specifying the glucose concentration by colour-coding these data points (that is, each data point now has three coordinates: uptake rate, extracellular glucose concentration, growth rate) causes a notable pattern to emerge (Fig. 3b). This analysis shows that growth rate $\mu$ is determined by two independent variables: the glucose uptake rate $r$, and the extracellular glucose concentration $g$. Our full experimental data set of all five single-*HXT* strains over a wide range of glucose and doxycycline concentrations are described by a single equation

$$\mu(r,g) = P(g) \times \ln\left(\frac{r}{r_c}\right) + \mu_c \qquad (1)$$

in which $\mu_c$ and $r_c$ are constants specifying the point of convergence of the log-linear lines (Fig. 3b), and the function $P(g)$ describes the slope of the log-linear correlation between $\mu$ and $r$ for each value of $g$. This equation does not depend on which Hxt transporter the cell uses for glucose uptake. This slope $P(g)$ increases with increasing $g$, and in turn tends to decrease growth rate (when $r < r_c$). $P(g)$ quantifies the marked effect that the extracellular glucose has on growth rate independently of glucose import—the effect of glucose perception. Qualitatively, equation (1) states that an increase in the extracellular glucose concentration may cause two counteracting effects: an increased glucose uptake rate $r$ (which tends to increase growth rate), and an increased perception of extracellular glucose (which tends to decrease growth rate). The net result on growth rate (that is, whether it rises or falls) is decided by the competition between these opposing effects of glucose perception and uptake. Which one of the two effects dominates depends on the actual values of $g$ and $r$, in particular on the product $P(g)\ln(r/r_c)$ quantifying the interaction between glucose perception and import (Supplementary Text).

The 'growth landscape' in Fig. 3c, described by equation (1), shows the full set of growth rates possible for a wide range of $g$ and $r$. Because equation (1) does not distinguish between the type and number of Hxt cells use for glucose import, it is applicable to cells with any number of *HXT* genes, including the wild-type, as long as the cells achieve the uptake rate within the range we probed. The shape of this landscape allows for the unusual growth-rate behaviours observed,

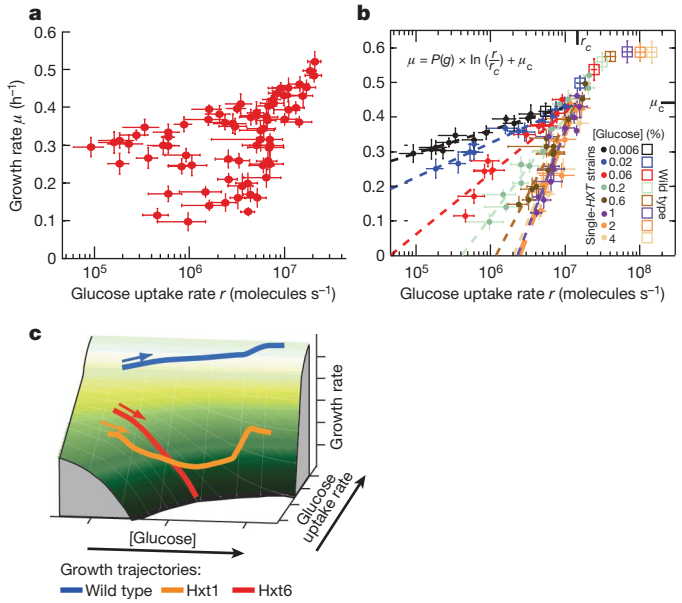



**Figure 3 | Emergence of a concise growth model incorporating cell's perception and uptake rate of glucose, and the resulting growth landscape.** **a**, **b**, Plotting together all of the measured growth rates and glucose uptake rates of the fluorescent single-*HXT* strains (**a**) then colour-coding by extracellular glucose level reveals this notable pattern (**b**). This plot shows that extracellular glucose concentration *g*, and glucose uptake rate *r*, are two independent variables. Growth rate is concisely described by the fit function $\mu(r, g)$. $P(g)$ is the slope of the log-linear correlation between growth rate and uptake rate for each *g*; we obtain $P(g)$ by fitting. $\mu_c$ and $r_c$ are constants specifying the point of convergence of the log-linear lines ($\mu_c = 0.44\ h^{-1}$, $r_c = 1.4 \times 10^7$ molecules $s^{-1}$). Error bars, s.e.m.; $n = 3$. **c**, Full growth landscape of budding yeast: Three-dimensional plot of the function $\mu(r, g)$. The growth trajectories followed by the parental wild-type (blue path, near the peak of this landscape), and fluorescent Hxt1-only and Hxt6-only strains (orange and red paths, respectively) are shown. Coloured arrows indicate the direction the cell travels on each path as *g* increases. The arrows along the two axes (along '[glucose]' and 'glucose uptake rate') point in the direction of increase.

including the convex-shaped growth rate of the Hxt1-only strain (Fig. 3c, orange path), the Hxt6-only strain's path towards growth arrest (Fig. 3c, red path), and the wild-type's hyperbolic growth rate (Fig. 3c, blue path). The wild-type strain is near the peak of this growth landscape yet its uptake rate is not much higher than that achieved by some single-*HXT* strains. The growth landscape shows that some values of (*g*, *r*) cannot sustain growth ($\mu = 0$). Indeed, for every *g*, there is a minimum uptake rate a cell needs to have for it to have any chance of growing in that particular glucose environment (Supplementary Fig. 9).

## Manipulation of glucose perception by sensors

Whereas the glucose uptake rate depends on the Hxt transporters, glucose perception, captured by $P(g)$, should depend on mechanisms the cell uses to measure the level of extracellular glucose. Snf3 and Rgt2 are two glucose sensors primarily known for regulating transcription of both major and minor glucose transporter genes[18,27] (*HXT*s, *GAL2*, *STL1* and *AGT1*). Because such regulation is disabled in our single-*HXT* strains, we could manipulate $P(g)$ by knocking out these two glucose sensors without affecting the uptake rate *r*. We constructed a panel of single-*HXT* strains with these two sensors deleted (Supplementary Fig. 10). The relationship between growth rates and extracellular glucose concentration in these 'sensorless' strains is notably different from that in strains with the two sensors intact (Fig. 4a and Supplementary Fig. 11). Growth rates now generally increase as the glucose level increases (at constant doxycycline level). Moreover, without the sensors, the Hxt3-only and Hxt6-only strains no longer approach growth arrest as the glucose level increases. Because we deleted all minor glucose transporter genes and removed the glucose's control of the sole transporter expression in our single-*HXT* strains, changes in uptake rate were not the reason for the growth rescues we observed. For every combination of glucose and doxycycline concentrations, the uptake rate of the sensorless strains was nearly identical to that of their sensor-containing counterparts (Fig. 4b and Supplementary Fig. 12).

In the sensorless strains, growth rate again explicitly depends on glucose concentration but with much reduced sensitivity (Fig. 4c, d). When Snf3 and Rgt2 are absent, a cell in 4% glucose acts as if it were in 0.06% glucose with intact sensors. Because the uptake rate remains virtually unchanged in the single-*HXT* strains when *SNF3* and *RGT2* are deleted, this reduced-sensitivity effect is due to a change in the perception function $P(g)$, not the uptake rate *r* (Fig. 4d). The remaining dependence of the cell's growth rate on the glucose concentration even after Snf3 and Rgt2 have been deleted suggests that other sensors may contribute to the effect embodied in $P(g)$[28,29]. Nonetheless, our experiments show that Snf3 and Rgt2 are the key determinants of $P(g)$ (as quantified in Fig. 4d).

The behaviour depicted by equation (1) should apply to the wild-type strain as well, as long as it achieves an uptake rate within the range probed with the single-*HXT* strains used to construct our growth landscape. We measured the wild-type's uptake rate and found that it was below the critical uptake rate $r_c$ for glucose concentrations smaller than 0.02% (Fig. 3b and Supplementary Fig. 16). For higher [glucose], the uptake rate exceeds $r_c$. When the wild-type cell's uptake rate is below $r_c$, its growth rate fits with the trend shown in Fig. 3b. For higher glucose concentration, the effect of perception on the wild-type's growth rate disappears (Fig. 3b). One possible explanation is that as long as the glucose concentration is not too low, the wild-type escapes the seemingly detrimental effect of perception on growth rate by making enough hexose transporters to go beyond $r_c$. But for lower glucose level in which its uptake is less than $r_c$, it properly tunes the interaction between glucose perception and uptake (quantified by the product $P(g)ln(r/r_c)$) such that its growth rate will increase when the cell perceives more extracellular glucose. Such tuning suggests that the transcriptional regulation of the *HXT* genes by Snf3 and Rgt2 is organized so that the wild-type always climbs uphill in the growth landscape (Fig. 3c) as it perceives an increase in the extracellular glucose concentration.

The critical point ($\mu_c$, $r_c$) may represent a region of phase transition in the cell's growth and metabolism. The cell markedly increases its ethanol production rate as its uptake rate increases above the critical rate $r_c$ (Supplementary Fig. 17). This suggests that when its uptake rate is below $r_c$, the cell metabolizes glucose mainly by respiration, but then switches to a largely fermentative metabolism as the uptake rate exceeds $r_c$. A key rate limiting step in fermentation is the import of glucose, and therefore the cell only redirects its glucose flux from respiration to fermentation when its glucose uptake rate is sufficiently high[30,31]. Our results indicate that this major redistribution of flux occurs around $r_c$.

## Discussion

Glucose perception and import are two separable modules that each affect the growth rate, but it is the interaction between them that ultimately determines the cell's growth rate, and that interaction can be both precisely altered and measured. The question remains as to why it would make sense for yeast to grow according to equation (1), which allows for a possible detrimental growth if the interaction between the perception and import modules is not properly tuned. One explanation may be that yeast has no way to directly 'measure' its glucose import rate in real-time. Indeed, there is no known 'flux sensor' that yeast uses to measure its glucose import rate in real-time and then adjust the production level of Hxt transporters to change the glucose import rate if it senses that the flux is too low. In fact, Hxt expression levels are primarily set by the extracellular glucose concentration[32] (Supplementary Fig. 1).

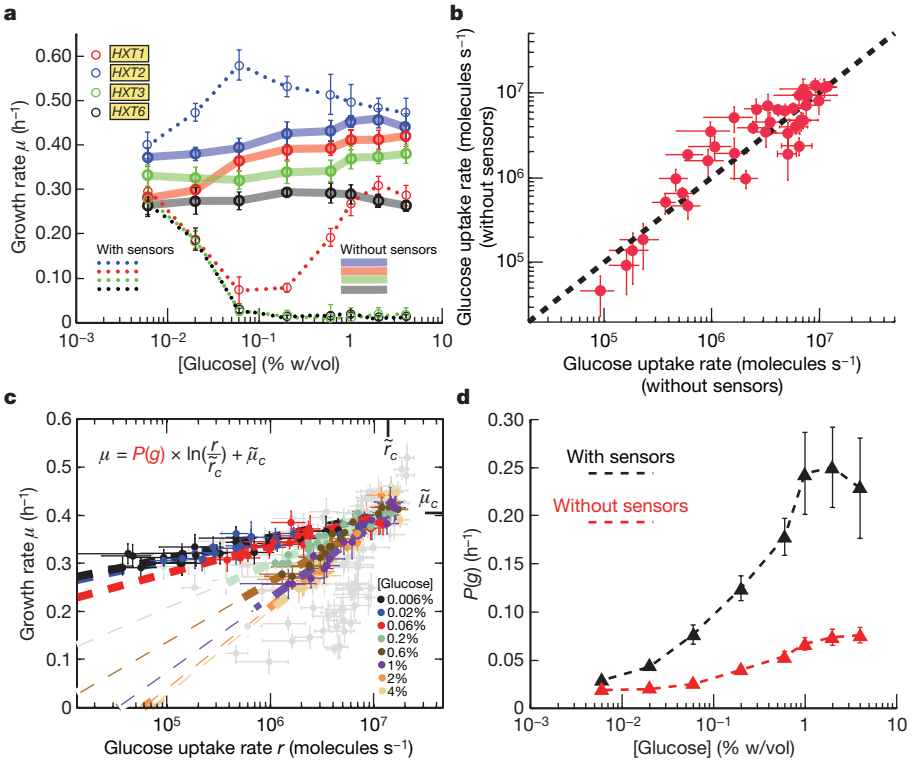



**Figure 4 | Manipulation of the cell's perception of extracellular glucose, leaving uptake rate unperturbed, can yield significant growth-rate changes.** **a**, Growth rates of single-*HXT* strains lacking two glucose sensors (*snf3*Δ *rgt2*Δ, bold lines) along with their counterparts with intact sensors (dotted lines) are shown for [doxycycline] = 5 μg ml$^{-1}$. Error bars, s.e.m.; $n = 3$. **b**, Knocking out the two glucose sensors leaves the cell's glucose uptake rate virtually unperturbed. Just the Hxt1-only and Hxt2-only strains are shown here for simplicity (see Supplementary Figs 7 and 12 for others). Each data point represents a particular combination of glucose and doxycycline concentrations. Error bars, s.e.m.; $n = 3$. **c**, By yEGFP fusion, fluorescent sensor-less single-*HXT* strains were constructed for comparison with their sensor-intact counterparts. The features of growth rates in **a** were preserved after this fusion (Supplementary Fig. 11). Growth rates and glucose uptake rates of these strains were measured (Supplementary Figs 12–15). For comparison, data for the sensor-intact single-*HXT* strains (from Fig. 3a) are shown in grey ($\tilde{\mu}_c = 0.40$ h$^{-1}$, $\tilde{r}_c = 1.4 \times 10^7$ molecules s$^{-1}$). Error bars, s.e.m.; $n = 3$. **d**, The sensitivity function $P(g)$, calculated from fitting the data in **c** and Fig. 3a is shown for strains with intact sensors (black) and *snf3*Δ *rgt2*Δ strains (red). Error bars indicate 95% confidence interval in these fits.

Although yeast certainly can measure the extracellular glucose level directly and the intracellular glucose level indirectly (for example, through the catabolite-repressor Mig1 that uses intracellular glucose as its substrate)[33–35], knowing the two glucose levels is not sufficient for yeast to infer what its glucose import rate is. This is because a given steady-state glucose concentration gradient can be maintained by a combination of wide ranges of glucose import rate and intracellular glucose breakdown rate. Because the cell has no direct way to measure the breakdown rate (there is no known 'rate sensor' measuring intracellular glucose breakdown), the cell cannot infer what the glucose import rate is in real-time just from the difference between extracellular and intracellular glucose. Given the engineering difficulty of building flux sensors, yeast may have solved the problem by evolving glucose sensors such as Snf3 and Rgt2 to measure the extracellular glucose level, then anticipate a certain glucose import rate would be achieved, set up intracellular activities to process glucose being imported at the anticipated rate, and make sure that such an import rate is indeed achieved by putting its *HXT* genes under the control of those glucose sensors (Supplementary Fig. 1).

Continuing efforts at large-scale modelling of glucose metabolism, gene regulation[36] and cellular signalling must decouple and consider how the cell's response varies when glucose uptake rate is varied independently of extracellular glucose level. For instance, microarray studies have shown that hundreds of genes involved in ribosomal biogenesis, which are energetically very costly, are upregulated several fold as yeast are subjected to ever increasing levels of glucose[37]. In these studies, as the level of glucose is increased, so does the glucose import rate. These observed large-scale changes are thus due to the conflated effects of glucose perception and import. It would be interesting to measure which of these changes are due to glucose perception and import separately by decoupling the two effects. We hope that our model, as well as the framework used to extract some key principles from the complexity underlying yeast growth, will assist continuing efforts to rationally engineer[38–40] and understand microbial metabolism at the systems level[41–48].

## METHODS SUMMARY

**Growth rate measurements.** Growth rates were measured while the cells were in log-phase growth in 5 ml batch cultures at 30 °C using synthetic media supplemented with the desired doxycycline and glucose concentrations. These concentrations remained nearly constant during growth (Supplementary Information). Using a spectrophotometer (Hitachi U-1800), we measured the absorbance at 600 nm ($A_{600\,nm}$) of these batch cultures over time, and extracted the growth rate of the cells.

**Glucose uptake rate measurements and calculations.** Glucose uptake rates were determined by measuring the rate of glucose depletion in the growth medium while the cells were in log-phase growth. It can be shown (Supplementary Information) that the glucose uptake rate per population density of cells (in units: mM h$^{-1}$ $A_{600\,nm}{}^{-1}$) is approximately $r(G_0) \approx \mu \frac{(G_0 - G(t))}{\rho(t) - \rho_0}$, in which $\rho(t) - \rho_0$ is the measured change in $A_{600\,nm}$ of the cell culture after time $t$, $\mu$ is the log-phase growth rate, and $G_0 - G(t)$ is the depleted glucose concentration in the growth medium after time $t$. This depleted glucose concentration was measured using a standard commercial glucose assay kit (Sigma) that is based on the conversion of glucose by hexokinase and $NADP^+$-dependent glucose-6-phosphate-dehydrogenase. We compared the measured glucose uptake rates with the uptake rates calculated using an independent method for the fluorescent single-*HXT* and wild-type strains. We calculated the glucose uptake rates by using the known Michaelis–Menten parameters ($V_{max}$ and $K_m$) of Hxt transporters[26] and the relative number of Hxt proteins per cell determined by measuring the average single-cell yEGFP fluorescence (Supplementary Information). These

comparisons showed a close agreement between our measured and calculated uptake rates (Supplementary Figs 6 and 14).

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 1 July; accepted 9 November 2009.**


1. Monod, J. *Recherches sur la Croissance des Cultures Bacteriennes* (Hermann et Cie, 1942).
2. Bennett, M. R. *et al.* Metabolic gene regulation in a dynamically changing environment. *Nature* **454,** 1119–1122 (2008).
3. Zaslaver, A. *et al.* Just-in-time transcription program in metabolic pathways. *Nature Genet.* **36,** 486–491 (2004).
4. Airoldi, E. *et al.* Predicting cellular growth from gene expression signatures. *PLoS Comp. Biol.* **5,** e1000257 (2009).
5. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436,** 588–592 (2005).
6. Krishna, S., Semssey, S. & Sneppen, K. Combinatorics of feedback in cellular uptake and metabolism of small molecules. *Proc. Natl Acad. Sci. USA* **104,** 20815–20819 (2007).
7. Ihmels, J., Levy, R. & Barkai, N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae. Nature Biotechnol.* **22,** 86–92 (2003).
8. Famili, I., Forster, J., Nielsen, J. & Palsson, B. O. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl Acad. Sci. USA* **100,** 13134–13139 (2003).
9. Bilu, Y., Shlomi, T., Barkai, N., & Ruppin, E. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS Comp. Biol.* 2, e106.
10. Levine, E. & Hwa, T. Stochastic fluctuations in metabolic pathways. *Proc. Natl Acad. Sci. USA* **104,** 9224–9229 (2007).
11. Fell, D. A. *Understanding the Control of Metabolism* (Portland, 1997).
12. Savageau, M. A. *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology* (Addison-Wesely, 1976).
13. Goyal, S. & Wingreen, N. S. Growth-induced instability in metabolic networks. *Phys. Rev. Lett.* **98,** 138105 (2007).
14. Nielsen, J., Villadsen, J. & Liden, G. *Bioreaction Engineering Principles* 235–311 (Springer, 2003).
15. Dickinson, J. R. & Schweizer, M. *The Metabolism and Molecular Physiology of Saccharomyces cerevisiae* (CRC, 2004).
16. Alon, U. Simplicity in biology. *Nature* **446,** 497 (2007).
17. Mallavarapu, A., Thomson, M., Ullian, B. & Gunawardena, J. Programming with models: modularity and abstraction provide powerful capabilities for systems biology. *J. R. Soc. Interface* **6,** 257–270 (2009).
18. Moriya, H. & Johnston, M. Glucose sensing and signaling in *Saccharomyces cerevisiae* through the Rgt2 glucose sensor and casein kinase I. *Proc. Natl Acad. Sci. USA* **101,** 1572–1577 (2004).
19. Boles, E. & Hollenberg, C. P. The molecular genetics of hexose transport in yeasts. *FEMS Microbiol. Rev.* **21,** 85–111 (1997).
20. Reifenberger, E., Freidel, K. & Ciriacy, M. Identification of novel *HXT* genes in *Saccharomyces cerevisiae* reveals the impact of individual hexose transporters on glycolytic flux. *Mol. Microbiol.* **16,** 157–167 (1995).
21. Bisson, L. F., Coons, D. M., Kruckeberg, A. L. & Lewis, D. A. Yeast sugar transporters. *Crit. Rev. Biochem. Mol. Biol.* **28,** 259–308 (1993).
22. Ozcan, S. & Johnston, M. Three different regulatory mechanisms enable yeast hexose transporter (*HXT*) genes to be induced by different levels of glucose. *Microbiol. Mol. Biol. Rev.* **63,** 554–569 (1999).
23. Pao, S. S., Paulsen, I. T. & Saier, M. H. Jr. Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* **62,** 1–34 (1998).
24. Wieczorke, R. *et al.* Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in *Saccharomyces cerevisiae. FEBS Lett.* **464,** 123–128 (1999).
25. Reifenberger, E., Boles, E. & Ciriacy, M. Kinetic characterization of individual hexose transporters of *Saccharomyces cerevisiae* reveals the impact of individual hexose transporters on glycolytic flux. *Eur. J. Biochem.* **245,** 324–333 (1997).
26. Maier, A., Volker, B., Boles, E. & Fuhrmann, G. F. Characterisation of glucose transport in *Saccharomyces cerevisiae* with plasma membrane vesicles (countertransport) and intact cells (initial uptake) with single Hxt1, Hxt2, Hxt3, Hxt4, Hxt6, Hxt7 or Gal2 transporters. *FEMS Yeast Res.* **2,** 539–550 (2002).
27. Walsh, M. C., Scholte, M., Valkier, J., Smits, H. P. & van Dam, K. Glucose sensing and signaling properties in *Saccharomyces cerevisiae* require the presence of at least two members of the glucose transporter family. *J. Bacteriol.* **170,** 2593–2597 (1996).
28. Jiang, Y., Davis, C. & Broach, J. Efficient transition to growth on fermentable carbon sources in *Saccharomyces cerevisiae* requires signaling through the Ras pathway. *EMBO J.* **17,** 6942–6951 (1998).
29. Boer, V. M., Amini, S. & Botstein, D. Influence of genotype and nutrition on survival and metabolism of starving yeast. *Proc. Natl Acad. Sci. USA* **105,** 6930–6935 (2008).
30. van Hoek, P., van Dijken, J. & Pronk, J. Effects of specific growth rate on fermentative capacity of baker's yeast. *Appl. Environ. Microbiol.* **64,** 4226–4233 (1998).
31. Reijenga, K. A. *et al.* Control of glycolytic dynamics by hexose transport in *Saccharomyces cerevisiae. Biophys. J.* **80,** 626–634 (2001).
32. Kaniak, A., Xue, Z., Macool, D., Kim, J. H. & Johnston, M. Regulatory network connecting two glucose signal transduction pathways in *Saccharomyces cerevisiae. Eukaryot. Cell* **3,** 221–231 (2004).
33. Gancedo, J. M. The early steps of glucose signaling in yeast. *FEMS Microbiol. Rev.* **32,** 673–704 (2008).
34. Kim, J. H. & Johnston, M. Two glucose-sensing pathways converge on Rgt1 to regulate expression of glucose transporter genes in *Saccharomyces cerevsiae. J. Biol. Chem.* **281,** 26144–26149 (2006).
35. Santangelo, G. M. Glucose signaling in *Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev.* **70,** 253–282 (2006).
36. Levy, S. *et al.* Strategy of transcription regulation in the budding yeast. *PLoS One* **2,** e250 (2007).
37. Yin, Z. *et al.* Glucose triggers different global responses in yeast, depending on the strength of the signal, and transiently stabilizes ribosomal protein mRNAs. *Mol. Microbiol.* **48,** 713–724 (2003).
38. Stephanopoulos, G. Challenges in engineering microbes for biofuels production. *Science* **315,** 801–804 (2007).
39. Lorenz, D. R., Cantor, C. R. & Collins, J. J. A network biology approach to aging in yeast. *Proc. Natl Acad. Sci. USA* **106,** 1145–1150 (2009).
40. Ostergaard, S., Olsson, L. & Nielsen, J. Metabolic engineering of *Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev.* **64,** 34–50 (2000).
41. Kell, D. B. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.* **7,** 296–307 (2004).
42. Savageau, M. A., Coelho, P., Fasani, R., Tolla, D. & Salvador, A. Phenotypes and tolerances in the design space of biochemical systems. *Proc. Natl Acad. Sci. USA* **106,** 6435–6440 (2009).
43. Ihmels, J. *et al.* Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309,** 938–940 (2005).
44. Klumpp, S. & Hwa, T. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proc. Natl Acad. Sci. USA* **105,** 20245–20250 (2008).
45. Duarte, N. C., Palsson, B., Ø. & Fu, P. Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae. BMC Genomics* **5,** 63 (2004).
46. Daran-Lapujade, P. *et al.* The fluxes through glycolytic enzymes in *Saccharomyces cereivisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl Acad. Sci. USA* **104,** 15753–15758 (2007).
47. Castrillo, J. I. *et al.* Growth control of the eukaryote cell: a systems biology study in yeast. *J. Biol.* **6,** 4 (2007).
48. Stelling, J. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.* **7,** 513–518 (2004).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank E. Boles for the kind gift of strains. We also thank D. Botstein, D. Muzzey, J. Gore and S. Rifkin for critical reading of our manuscript and useful discussions. This work was funded by a National Institutes of Health (NIH) Director's Pioneer awarded to A.v.O., and grants from the NIH and National Science Foundation (NSF). H.Y. was supported by the Natural Sciences and Engineering Research Council of Canada's (NSERC) Graduate Fellowship.

**Author Contributions** H.Y. performed the experiments. H.Y. and A.v.O. designed experiments, analysed data and wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.v.O. (avano@mit.edu).

## METHODS

**Strain background and construction.** A list of strains with diagrams summarizing their key features is provided in the Supplementary Information. All strains were derived from the haploid strain CEN.PK2-1C (*MATα*, gift from E. Boles)[24], referred to as the 'wild-type' in our study. Both EBY.VW4000 and EBY.VW5000 are deficient in hexose transport owing to deletions of all *HXT* genes as well as genes encoding transporters with minor glucose uptake capabilities (*agt1*Δ *ydl247w*Δ *yjr160*Δ)[24]. HY4D1 and HY5F1 each contain reverse tetracycline-controlled transactivator (rtTA) protein expressed constitutively by the *MYO2* promoter (inserted into EBY.VW4000 and EBY.VW5000 respectively using plasmid pDH18 (EUROSCARF) containing *HIS5* gene) and CFP constitutively expressed by $P_{TEF1}$. XhoI-$P_{TET07}$-BamHI, BamHI-HXTn-NotI fragments were cloned into pRS305 (EUROSCARF) backbone containing the *LEU2* gene ($n = 1$–4, 6). Integrating these plasmids into the defective *LEU2* locus (leu2–3) in HY4D1 by linearizing the plasmids with NarI, the single-*HXT* strains were constructed. To construct fluorescent single-*HXT* strains, the yEGFP-$T_{ADH1}$-Kan fragment was amplified from the pKT127 plasmid (EUROSCARF) and fused to the carboxy terminus of *HXTn* open-reading frame (ORF) in each of the single-*HXT* strains by standard PCR integration[49]. This fragment was also fused to C terminus of *HXTn* ORF ($n = 1$–4, 6, 7) in CEN.PK2-1C, thus resulting in six fluorescent wild-type strains (Supplementary Fig. 16). The 'sensorless' versions of single-*HXT* strains (*snf3*Δ *rgt2*Δ) were constructed in the same way as their sensor-intact counterparts mentioned earlier, by using HY5F1 instead of HY4D1. To probe the wild-type's transcriptional regulation of each of the *HXT* genes (Supplementary Fig. 1), XhoI-$P_{HXTn}$-BamHI, BamHI-YFP-NotI fragments were cloned into pRS305 backbone containing the *LEU2* gene ($n = 1$–4, 7) and integrated into the defective *LEU2* locus (leu2–3) in CEN.PK2-1C by linearizing the plasmid with either NarI (for $n = 1$) or ClaI (for all other $n$), resulting in five strains. The $P_{HXT1}$, $P_{HXT2}$, $P_{HXT3}$, $P_{HXT4}$ and $P_{HXT7}$ promoter sequences refer to 1,941-, 850-, 1,996-, 1,544- and 2,042-base pairs upstream of the start codon of the respective genes. These sequences include all the known binding sites of transcription factors for the respective genes[50].

**Growth rate measurements.** All growth rates reported in our study were measured while the cells were in log-phase growth in 5 ml batch cultures at 30 °C, in a standard synthetic media with various combinations of glucose and doxycycline concentrations. To bring the cells into log-phase, the single-*HXT* strains were first grown in a standard synthetic media containing 2% maltose and the desired concentration of doxycycline until the cells have been in log-phase for roughly 12 h. This procedure ensured that the cells were already making Hxt proteins needed to initiate glucose uptake immediately after being transferred to glucose media. Then these cells were diluted into the standard synthetic media with the same amount of doxycycline, but this time containing glucose instead of maltose. These dilutions were done such that by the time the density of cells in the batch culture reached a level detectable by our spectrophotometer (Hitachi U-1800) (roughly 15 h after dilution), the cells had adjusted to the glucose media and were in log-phase growth. Hence, the transient growth rate change associated with maltose to glucose media transfer did not enter into our growth rate measurements. In a separate experiment, we confirmed this was indeed the case by further diluting these cultures into an identical glucose media, which showed that having the cells pre-grown in maltose before did not affect the growth rates reported in our study. By measuring the $A_{600\,nm}$ of these batch cultures over time, we extracted the growth rate of the cells. Strains that approached growth arrest also went through the same procedure as above. After transfer from maltose to glucose media, these cells' growth rates transiently decreased to nearly zero during a period of roughly 24 h. By looking at the cells under the microscope, no abnormal cell morphologies were detected, thus indicating normal growth (that is, no pseudohyphal or filamentous growth was detected).

**Fluorescence measurements.** The average single-cell fluorescence due to yEGFP fused to the C terminus of *HXT* genes in the wild-type and single-*HXT* strains was measured using a Becton Dickinson FACScan flow cytometer with excitation laser at 488 nm. Emission filter FL1 (530/30) was used to detect the yEGFP fluorescence levels as well as the YFP for determining the $P_{TET07}$ induction curves in the calibration strains HY4DCal5 and HY5FCal2. Before observation using FACScan, the strains were grown using the protocol outlined in the 'growth rate measurements' section. The mean fluorescence values reported in our study represent the steady-state levels of Hxt proteins in single cells, as no appreciable changes in fluorescence was detected while the cells were growing in log-phase.

**Glucose uptake rate measurements and calculations.** Glucose uptake rates of cells were determined by measuring the rate of glucose depletion in the growth medium while the cells were in log-phase growth. First, the reasoning behind this procedure is as follows: If the cell's growth rate at glucose concentration $G_0$ is $\mu$, $G(t)$ is the concentration of glucose in the growth medium at time $t$, $r(G(t))$ is the uptake rate per absorbance of the cells as a function of extracellular glucose, and $r_0$ is the absorbance of cells at $t = 0$, then the decrease in glucose concentration in the growth medium over time $t$ is

$$G_0 - G(t) = \int_0^t r(G(\tau))\rho_0 \exp(\mu\tau)d\tau \quad (2)$$

If this change in glucose concentration is sufficiently small, but large enough to be detectable by our chemical assay (described below), then we can approximate $r(G(t)) \approx r(G_0)$ and $\mu$ as a constant during the time interval $t$. Then equation (2) can be solved for $r(G_0)$:

$$r(G_0) \approx \mu \frac{(G_0 - G(t))}{\rho(t) - \rho_0} \quad (3)$$

in which $r(G_0)$ is the uptake rate per $A_{600\,nm}$, measured in units of $mM\,h^{-1}\,A_{600\,nm}{}^{-1}$. This was then converted into molecules $s^{-1}$ $cell^{-1}$ using conversion factor $1.7 \times 10^7$ cells $ml^{-1}$ $A_{600\,nm}{}^{-1}$. $\rho(t) - \rho_0$ is the change in $A_{600\,nm}$ of the cells measured using the spectrophotometer (Hitachi U-1800), and $\mu$ is the growth rate determined by the method mentioned previously. The change in glucose concentration $G_0 - G(t)$ was measured using the standard commercial glucose assay kit (Sigma) based on conversion of glucose through hexokinase and $NADP^+$ dependent glucose-6-phosphate-dehydrogenase. We compared the measured glucose uptake rates with the uptake rates calculated using an independent method for the fluorescent single-*HXT* and wild-type strains. We calculated the glucose uptake rates by using the known Michaelis–Menten parameters ($V_{max}$ and $K_m$) of Hxts[26] and the relative number of Hxt proteins per cell inferred from measuring the average single-cell yEGFP fluorescence (Supplementary Information). These comparisons showed a close agreement between our measured and calculated uptake rates (Supplementary Figs 6 and 14).

49. Sheff, M. & Thorn, K. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast* **21**, 661–670 (2004).
50. Kim, J. H., Polish, J. & Johnston, M. Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor Rgt1. *Mol. Cell. Biol.* **23**, 5208–5216 (2003).

Vol 462|17 December 2009|doi:10.1038/nature08616

ARTICLES

# Transport mechanism of a bacterial homologue of glutamate transporters

Nicolas Reyes[1], Christopher Ginter[1] & Olga Boudker[1]

**Glutamate transporters are integral membrane proteins that catalyse a thermodynamically uphill uptake of the neurotransmitter glutamate from the synaptic cleft into the cytoplasm of glia and neuronal cells by harnessing the energy of pre-existing electrochemical gradients of ions. Crucial to the reaction is the conformational transition of the transporters between outward and inward facing states, in which the substrate binding sites are accessible from the extracellular space and the cytoplasm, respectively. Here we describe the crystal structure of a double cysteine mutant of a glutamate transporter homologue from *Pyrococcus horikoshii*, $Glt_{Ph}$, which is trapped in the inward facing state by cysteine crosslinking. Together with the previously determined crystal structures of $Glt_{Ph}$ in the outward facing state, the structure of the crosslinked mutant allows us to propose a molecular mechanism by which $Glt_{Ph}$ and, by analogy, mammalian glutamate transporters mediate sodium-coupled substrate uptake.**

Glutamate is the predominant excitatory neurotransmitter in the brain responsible for learning, memory formation and higher cognitive function. Specialized membrane proteins—glutamate transporters (also termed excitatory amino acid transporters (EAATs))—mediate the transmitter uptake from the extracellular space into the cytoplasm of astrocytes and neurons against steep concentration gradients to allow for rounds of neurotransmission and to prevent glutamate-mediated excitotoxicity[1]. EAATs are members of a ubiquitous solute carrier 1 (SLC1) family of secondary solute transporters[2], which catalyse concentrative uptake of the acidic and neutral amino acids and dicarboxylic acids in a reaction coupled to symport of sodium and/or protons and, in the case of EAATs, antiport of potassium. The conceptual mechanisms of protein-mediated solute transport were developed over 40 years ago[3–5] and focus on the ability of transporters to undergo isomerization between two states: an outward facing state, in which the substrate-binding site is accessible from the external solution, and an inward facing state, in which it is reached from the cytoplasm. Although our understanding of the molecular mechanisms of transport have advanced significantly in the last decade[6,7], the structural transitions between the outward and inward facing states remain poorly characterized.

The sodium/aspartate symporter from *Pyrococcus horikoshii* ($Glt_{Ph}$), an archaeal homologue of the EAATs, was one of the first sodium-coupled transporters for which structure had been determined at a near-atomic resolution[8,9]. The structural analysis of $Glt_{Ph}$ revealed that individual protomers assemble into a homotrimer, forming a deep solvent-filled bowl open to the extracellular solution and reaching approximately half way across the lipid bilayer (Fig. 1a). The first six transmembrane (TM) segments of each $Glt_{Ph}$ protomer mediate all inter-subunit contacts and form a distorted cylinder within which the highly conserved carboxy terminus of the protein is folded into a compact core. The core consists of an intracellular re-entrant helical hairpin (HP) 1, TM7 with an unwound segment, an extracellular hairpin, HP2, and an amphipathic TM8. Despite the trimeric assembly, shared by all characterized glutamate transporter homologues[10–12], protomers function independently[13–17]. Consistently, in $Glt_{Ph}$ crystal structure an L-aspartate (L-Asp) molecule and two sodium ions are bound within the core of each individual protomer. Positioned between the tips of HP1 and HP2 at the bottom of the extracellular bowl, the substrate is occluded from the aqueous environment only by HP2. Notably, in a crystal structure of $Glt_{Ph}$ in complex with a competitive blocker, L-threo-β-benzyloxyaspartate (L-TBOA)[9,18], HP2 assumes an open conformation, revealing its dynamic nature and suggesting that it may serve as an extracellular gate. Bound L-Asp and L-TBOA are separated from the cytoplasm by over 15 Å of a compact protein core, indicating that these structures correspond to the outward facing state of the transporter. However, the structure of the inward facing state and the manner in which the substrate and sodium ions are released into the intracellular solution remain unknown.

## Crosslinking cysteines in TM2 and HP2

In a study predating the crystal structure determination of $Glt_{Ph}$, it was reported that two cysteines placed in TM2 and HP2 of EAAT1 formed a spontaneous disulphide bond, suggesting close proximity of these residues[19]. Notably, in the crystal structures of L-Asp-and L-TBOA-bound $Glt_{Ph}$ the corresponding residues K55 in TM2 and A364 in HP2 are over 25 Å apart and cysteines at these positions are unlikely to become crosslinked. Moreover, the inter-subunit residue-to-residue distances are also 25–30 Å (Fig. 1a). Intrigued by these results, we reasoned that TM2 and HP2 become juxtaposed in an as yet structurally uncharacterized functional state of the transporter and set out to reproduce the double cysteine mutation in $Glt_{Ph}$ for crystallographic studies. The K55C/A364C mutant was generated within a cysteine-less $Glt_{Ph}$, expressed in *Escherichia coli* and purified in detergent solution. Oxidative crosslinking of $Glt_{Ph}$(K55C/A364C) in the presence of copper 1,10-phenanthroline (CuPhen) yielded distinct protein species with a higher electrophoretic mobility (Fig. 1b), which we identified as intra-molecularly crosslinked $Glt_{Ph}$ protomers (Supplementary Fig. 1a–c). Interestingly, CuPhen-catalysed disulphide bond formation is greatly facilitated when the transporter is purified in the absence of sodium ions ($Na^+$) and L-Asp, suggesting that the state in which residues 55 and 364 are proximal is more populated under these conditions (Fig. 1b and Supplementary Fig. 2a). Qualitatively similar results were also obtained using unpurified K55C/A364C mutant within *E. coli* crude

[1]Department of Physiology and Biophysics, Weill Cornell Medical College, 1300 York Avenue, Box 75, New York, New York 10065, USA.

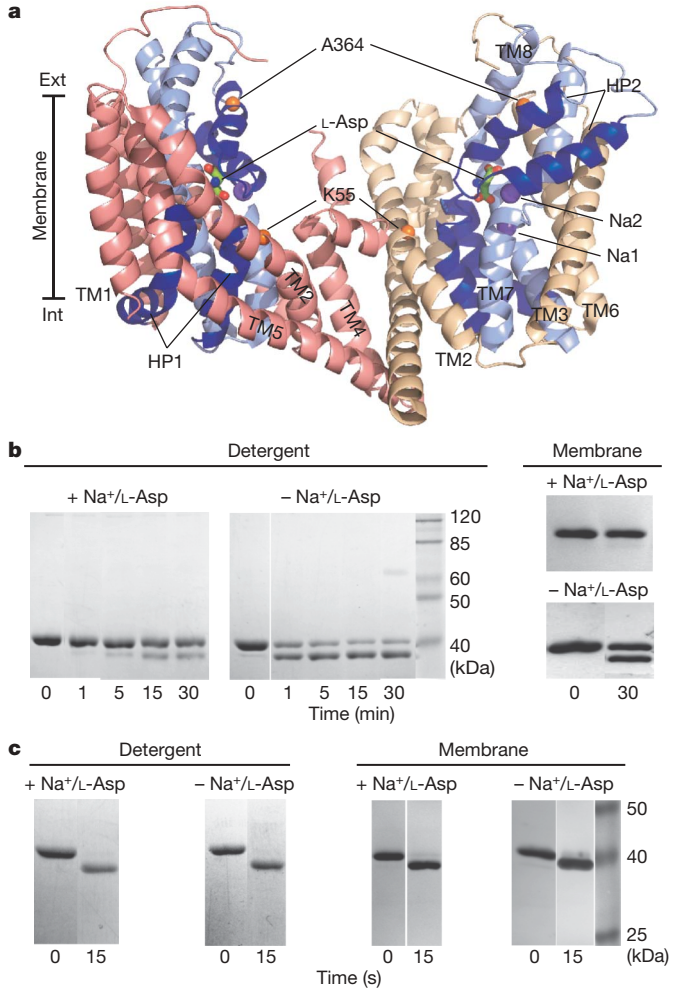



**Figure 1 | Crosslinking of $Glt_{Ph}$(K55C/A364C). a**, Cartoon representation of two substrate-bound $Glt_{Ph}$ protomers (Protein Data Bank accession 2NWX) viewed in the membrane plain. The third protomer and TM4 of the protomer on the right are removed for clarity. TM1–TM6 are coloured salmon and wheat; the C-terminal cores are light blue; and the hairpins HP1 and HP2 are dark blue. Bound L-Asp and $Na^+$ are shown as sticks and purple spheres, respectively. Orange spheres correspond to the Cα atoms of residues 55 and 364. All molecular representations have been generated using Pymol[46]. **b**, SDS–PAGE analysis of $Glt_{Ph}$(K55C/A364C) before and after incubation with 100 μM CuPhen for indicated periods of time. Detergent-solubilized purified $Glt_{Ph}$(K55C/A364C) (left) and unpurified transporter in crude *E. coli* membranes (right) were visualized by Coomassie staining and western blotting, respectively. **c**, Crosslinking of $Glt_{Ph}$(K55C/A364C) in the presence of 50 μM $HgCl_2$. Samples were analysed as in **b**.

membranes, demonstrating that $Glt_{Ph}$, like EAAT1, adopts the new conformation in the context of a lipid bilayer. Moreover, cysteines at structurally adjacent positions did not form a disulphide bond, suggesting that the distance shortening between positions 55 and 364 occurs in a highly specific manner (ref. 19 and Supplementary Fig. 2b). Notably, incubation of $Glt_{Ph}$(K55C/A364C) with divalent mercury ($Hg^{2+}$), which serves as a bi-functional thiol-specific crosslinker, yields an essentially complete crosslinking both in the presence and in the absence of $Na^+$ and L-asp in detergent and in crude membranes (Fig. 1c). Furthermore, crosslinking is completed within seconds, suggesting that both substrate-free and bound $Glt_{Ph}$ rapidly sample the crosslinking-competent state.

## Crystal structure of the crosslinked $Glt_{Ph}$(K55C/A364C)

$Hg^{2+}$-crosslinked K55C/A364C mutant ($Glt_{Ph}$(55C/$364C_{Hg}$)) in the presence of $Na^+$ and L-Asp yielded crystals diffracting to 3.5–3.9 Å (see Supplementary Table 1). A previous extensive study suggested that the functionally relevant conformational transitions of a related bacterial glutamate transporter do not involve regions engaged in protein trimerization, because crosslinking of multiple interfacial residues had no effect on substrate transport[17]. Hence, we used a trimeric $Glt_{Ph}$ model comprised of TM2, TM4 and TM5, which are responsible for all inter-subunit contacts (Fig. 1a), to search for a molecular replacement solution. The initial phases yielded distinct peaks within the anomalous difference Fourier map corresponding to Hg atoms at positions adjacent to cysteine 55 in TM2. Rounds of manual building and refinement resulted in a protein model lacking 5 and 9 residues on the amino and carboxy termini, respectively, and including 98% of the side chains. To verify the model, we crystallized the selenomethionine-substituted $Glt_{Ph}$(55C/$364C_{Hg}$). The Se peaks within the anomalous difference Fourier maps are in perfect agreement with the location of 17 methionines in the protein model (Supplementary Fig. 3a). The distance between the Cα atoms of modelled cysteines 55 and 364 is 7.4 Å, and the Hg peak within the anomalous difference Fourier map is positioned between these residues with the sulphur to mercury distances estimated at 2–2.3 Å (Supplementary Fig. 3b), similar to those observed in small molecular mass compounds[20] and peptides[21,22].

## $Glt_{Ph}$(55C/$364C_{Hg}$) is captured in the inward facing state

The most striking structural feature of $Glt_{Ph}$(K55C/$364C_{Hg}$) compared to the wild-type $Glt_{Ph}$ is an approximately 18 Å movement of the substrate binding sites from near the extracellular solution at the bottom of the extracellular bowl to near the cytoplasm (Fig. 2a, b). Structural superposition using transmembrane segments TM1, TM2, TM4 and TM5 from all three protomers reveals that these segments remain largely unchanged, with the root mean squared deviation (r.m.s.d.) for the atomic positions of 1.2 Å. The remainder of the protein, including TM3, TM6, HP1, TM7, HP2 and TM8, does not align and has undergone a substantial movement (Fig. 2a). Remarkably, when these regions alone are superposed, their structures are essentially identical, yielding r.m.s.d. of 0.6 Å. These results suggest that $Glt_{Ph}$ can be partitioned into two structural domains, which we termed the trimerization and transport domains (Fig. 2c). The rigid body movements of the transport domains relative to the trimerization domain dominate the structural changes between the substrate-bound wild-type $Glt_{Ph}$ and $Glt_{Ph}$(55C/$364C_{Hg}$) (Fig. 2a, b and Supplementary Movie). The structural environment of the ion- and substrate-binding sites, which are positioned entirely within the transport domains, is preserved. The $F_o - F_c$ difference Fourier map reveals excess electron density at these sites (Supplementary Fig. 4), suggesting that the transporter remains bound to L-Asp and to two $Na^+$ ions. Consistently, substrate-free $Glt_{Ph}$(55C/$364C_{Hg}$) in solution binds L-Asp with high affinity and in a sodium-dependent manner (Fig. 2d). In the structure of $Glt_{Ph}$(55C/$364C_{Hg}$) the substrate binding sites lie near the bottom of deep crevices formed within each protomer between HP1 and the cytoplasmic portion of TM7 of the transport domain and TM2 and TM5 of the trimerization domain (Fig. 2b). Bound L-Asp and Na1 are completely occluded from the solution by the tips of HP1 and HP2, whereas Na2 is partially exposed to the solvent (Fig. 2a, b). Because substrate and ion binding sites are near the intracellular solution, we propose that the conformation of $Glt_{Ph}$(55C/$364C_{Hg}$) corresponds to an inward facing state of the transporter.

Because the trimerization domain undergoes little conformational transition, it is unlikely to change its position relative to the membrane plane. Hence, we estimate that a ~25 Å thick apolar region of the lipid bilayer, which in the wild-type $Glt_{Ph}$ aligns approximately with the hydrophobic portion of TM1, is similarly positioned in $Glt_{Ph}$(55C/$364C_{Hg}$). Such bilayer placing suggests that several polar residues at the extracellular ends of the lipid-facing TM3 and TM6 move into the hydrophobic region of the membrane. It is possible that the lack of the lipid bilayer in the crystals removes important constraints, resulting in a non-native positioning of the transport domain. However, crosslinking of residues 55C and 364C in $Glt_{Ph}$

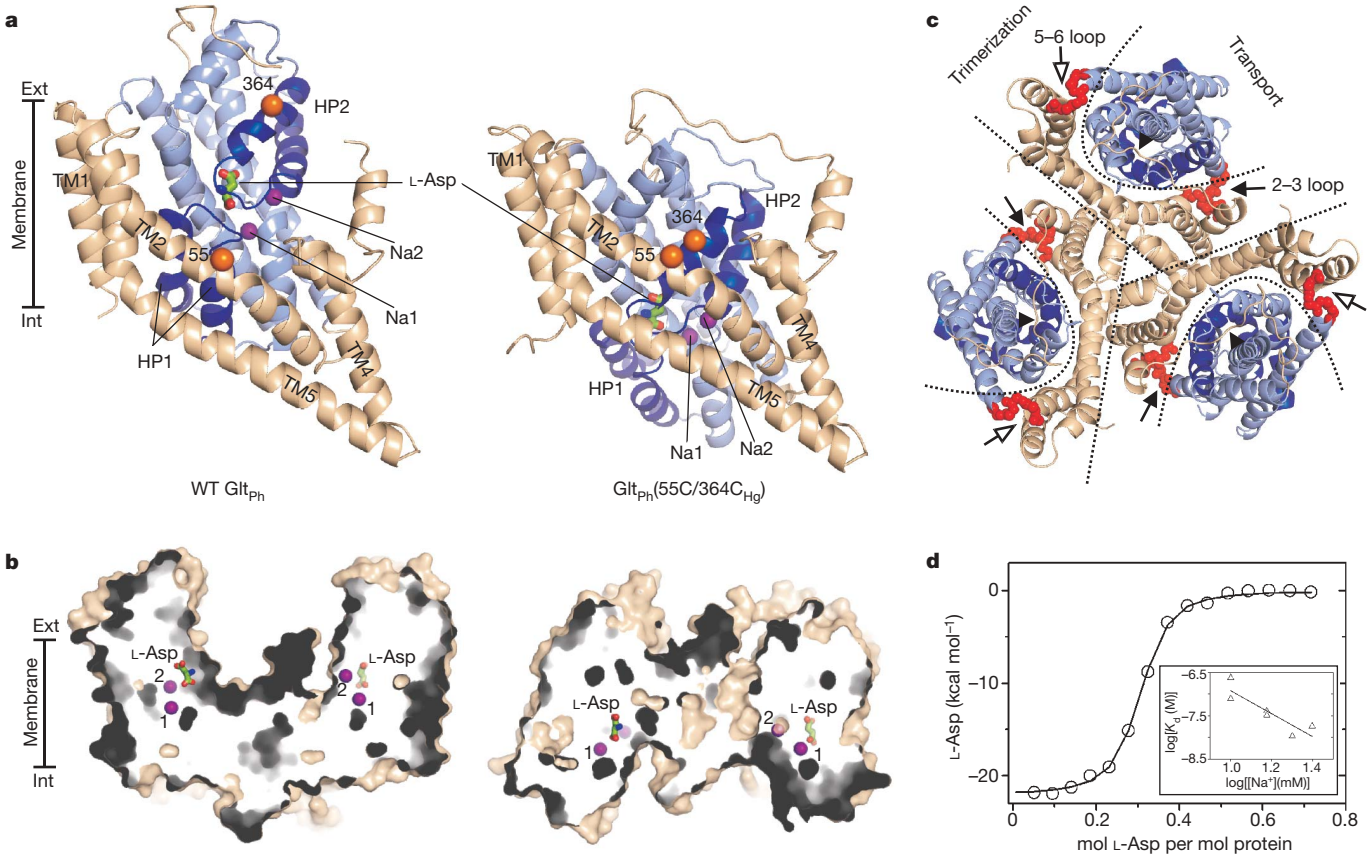



**Figure 2 | $Glt_{Ph}$(55C/364$C_{Hg}$) in the inward facing substrate-bound state.** **a**, **b**, Cartoon representation of the single protomers (**a**) and surface representation of the trimers sliced through the binding sites (**b**). Wild-type (WT) $Glt_{Ph}$ and $Glt_{Ph}$(55C/364$C_{Hg}$) are shown on the left and right, respectively. In **a**, TM1, TM2, TM4 and TM5 are coloured wheat, with the remainder of the protomer light blue. $Na^+$ ions are shown as purple spheres. **c**, Extracellular view of $Glt_{Ph}$(55C/364$C_{Hg}$) with straight and curved lines delineating individual protomers and transport domains, respectively. The trimerization domains (wheat) and transport domains (blue) are connected by short cytoplasmic (filled arrows) and extracellular (open arrows) loops, highlighted in red. The long TM3–TM4 loops (arrowhead) cross over the transport domains. **d**, Isothermal titration calorimetry analysis of L-Asp binding to $Glt_{Ph}$(55C/364$C_{Hg}$). Shown are the binding heats in 10 mM NaCl. The linear dependence (slope = $-2.6 \pm 0.7$) of the log of the apparent dissociation constant, $K_d$, on the log of $Na^+$ concentration is shown in the inset. $Glt_{Ph}$(55C/364$C_{Hg}$) was exchanged into $Na^+$/L-Asp-free buffer, diluted to 15–20 μM in the reaction cell of the Microcal ITC$^{200}$, supplemented with indicated concentrations of $Na^+$, and titrated with L-Asp at 25 °C. The binding enthalpy and the apparent number of binding sites were $23.6 \pm 0.8$ kcal $mol^{-1}$ and $0.4 \pm 0.03$ ($n = 6$), respectively.

and corresponding residues in EAAT1 occurs efficiently in lipid membranes, suggesting that a sufficient range of motions of the transport domain is allowed. Movement of the transport domain towards the cytoplasm also significantly increases its exposure to the intracellular solvent. Notably, although the total exposed cytoplasmic surface area increases from ~1,300 Å$^2$ in the wild type to ~4,200 Å$^2$ in $Glt_{Ph}$(55C/364$C_{Hg}$), the fraction of the apolar area changes only modestly from 57% to 63%. Both values are well within the range reported for the solvent-exposed area of soluble proteins[23].

## Domain interaction interfaces

Both wild-type $Glt_{Ph}$ and $Glt_{Ph}$(55C/364$C_{Hg}$) bury approximately equal surface areas of ~2,500 Å$^2$ on the interface between the trimerization and transport domains. Furthermore, in both structures essentially the same residues of the trimerization domain are involved, forming a smooth relatively featureless interaction surface, which is over 80% hydrophobic (Fig. 3a, b). Although TM1 and TM4 contribute to the contact area, TM2 and TM5, lying parallel to each other and crossing the membrane at an oblique angle, provide the bulk of the interactions, contributing 77% and 67% to the interaction surface in the wild type and $Glt_{Ph}$(55C/364$C_{Hg}$), respectively. In contrast, the transport domains in the two structures present interaction interfaces with little overlap, but similar in area (1,250 Å$^2$) and hydrophobicity (70%). They are dominated by HP1 and the cytoplasmic portion of TM7 in the wild-type $Glt_{Ph}$ (Fig. 3a), and by HP2 and the extracellular portion of TM8 in $Glt_{Ph}$(55C/364$C_{Hg}$) (Fig. 3b), which provide 92% and 88% of the total buried area, respectively. The analysis of the evolutionary conservation reveals that the interacting residues within the trimerization domain as well as within the two alternative interfaces of the transport domain are all highly conserved (Fig. 3c, d). The high conservation of HP2 surface residues and the reported transport inhibition of the mammalian EAATs upon chemical modification of cysteines in HP2 (refs 24–28) are consistent with this region being involved in specific protein contacts as observed in $Glt_{Ph}$(55C/364$C_{Hg}$). The disengagement of HP1 and the N-terminal part of TM7 and their transition to the cytoplasmic surface of the transporter is also consistent with the previously reported intracellular solvent accessibility of these regions in related prokaryotic and mammalian transporters[29–32].

It has been noted previously that the C-terminal core of $Glt_{Ph}$ contains structurally symmetrical elements[8]. Specifically, HP1 and the N-terminal half of TM7 can be superimposed on HP2 and the N-terminal portion of TM8 with r.m.s.d. of 2.5 Å. It is precisely these triplets of helices that constitute the two alternative interaction interfaces of the transport domain (Fig. 3a, b). Notably, also the N-terminal portion of the transporter, comprised of the first six transmembrane segments, exhibits a pseudo-two-fold structural symmetry despite the lack of detectable sequence conservation. TM4–TM6 can be superimposed upon TM1–TM3 with r.m.s.d. of ~6 and 4.3 Å for the wild type and mutant, respectively (Fig. 4a, b).

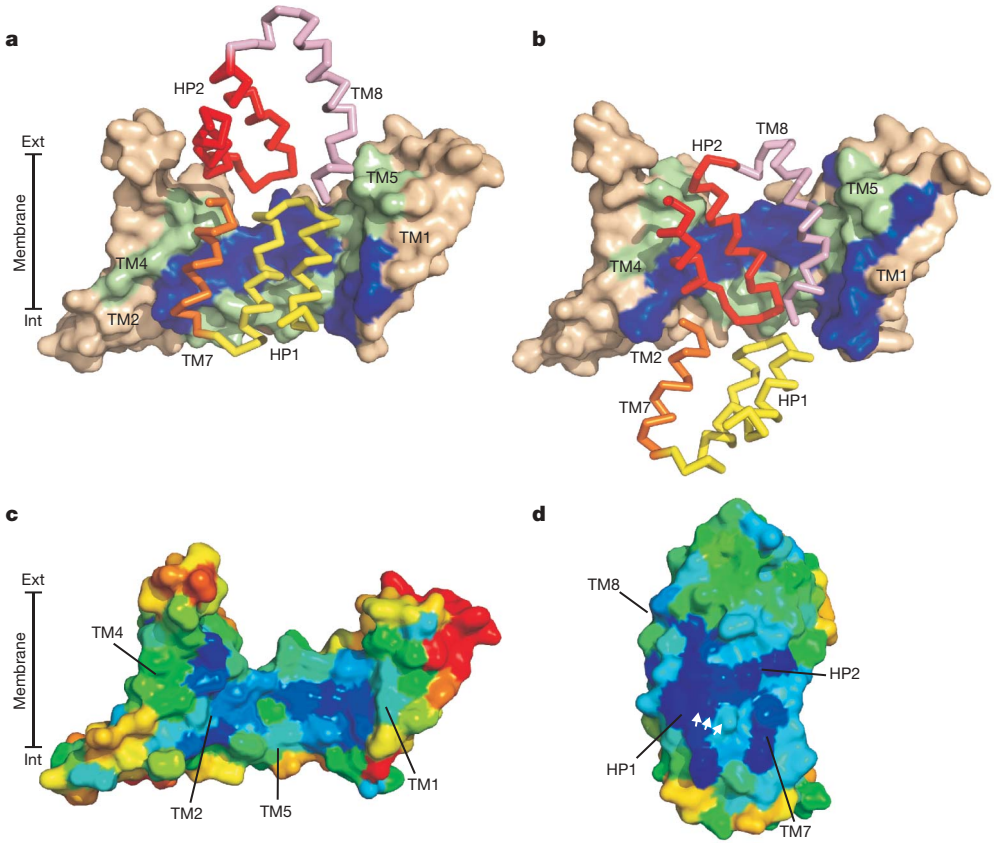



**Figure 3 | Domain interaction surfaces.** **a**, **b**, Trimerization domains of wild-type $Glt_{Ph}$ (**a**) and $Glt_{Ph}(55C/364C_{Hg})$ (**b**) are shown in surface representation and coloured wheat. Residues involved in domain contacts, identified by ProFace server[47], are coloured blue (TM1 and TM2) and green (TM4 and TM5). Interacting structural elements of the transport domain are shown in ribbon representation: HP1 (yellow) and TM7 (orange) in wild-type $Glt_{Ph}$; HP2 (red) and TM8 (pink) in $Glt_{Ph}(55C/364C_{Hg})$. **c**, **d**, Surface representation of the trimerization (**c**) and transport (**d**) domains coloured according to evolutionary conservation. Dark blue and red correspond to the highly conserved and variable residues, respectively. The interacting surfaces are facing the viewer, and the white arrows mark the highly conserved serine-rich signature motif in HP1. Conservation scores were calculated using the Consurf server[48], and 212 SLC1 sequences with less than 60% identity were harvested from the Pfam database[49] and aligned in ClustalW2[50].

Hence, unlike other characterized secondary transporters[33–35], glutamate transporters contain not one but two inverted structural repeats (Fig. 4d). Because of the N-terminal symmetry, the interaction interface of the trimerization domain can be partitioned into structurally related extracellular and cytoplasmic portions (Fig. 4c), which contribute approximately equally to the total buried surface area.

## Hinge movements

Within the transport domain, structurally symmetrical TM3 and TM6 serve as two arms holding the transporter core and extending from the trimerization domain: the former from the cytoplasm and the latter from the extracellular solution (Fig. 4e, f). The loops, which connect these transmembrane segments to the trimerization domain, enable the inward movement. The long flexible loop between TM3 and TM4 is poorly structured in the L-Asp-bound wild-type $Glt_{Ph}$ crystals, and in $Glt_{Ph}(55C/364C_{Hg})$ its structure is largely determined by crystal contacts. In contrast, the structurally symmetrical loops between TM2 and TM3 and between TM5 and TM6 are short and undergo well-defined transitions to accommodate the movements of the transport domain (Fig. 4e, f and Supplementary Movie). In $Glt_{Ph}(55C/364C_{Hg})$, unwinding of the TM3 α-helix by approximately one turn extends the TM2–TM3 loop by four residues and allows for the descent of TM3 towards the cytoplasm. In contrast, the TM5–TM6 loop is shortened by two residues as a result of simultaneous unwinding of TM5 by one-half of a turn and extension of TM6 by a turn of a helix. Both loops contain conserved glycines—G69, G79 and G82 in the TM2–TM3 loop and G221 in the TM5–TM6 loop—which may facilitate the folding/unfolding of the helices and serve as hinges. Indeed, the transport domain movement can be decomposed into a 16 Å descent towards the cytoplasm followed by a 37° rotation around an axis passing simultaneously through the transport domain centre of mass and loops TM2–TM3 and TM5–TM6. Consistent with the importance of the structural re-arrangements in the loop regions, chemical modifications of cysteines in TM5–TM6 loop of EAAT3 inhibited uptake[36].

## Transport mechanism

In the structure of wild-type $Glt_{Ph}$, L-Asp is bound ~5 Å beneath the surface of the extracellular bowl occluded from the solution by HP2. In contrast, in the structure of $Glt_{Ph}(55C/364C_{Hg})$, L-Asp is bound ~5 Å beneath the intracellular surface occluded from the solution by HP1. Hence, we propose that the structure of L-Asp-bound $Glt_{Ph}$ corresponds to the state right after binding of the substrate from the extracellular side (outward-facing occluded) and the $Glt_{Ph}(55C/364C_{Hg})$ structure corresponds to the state right before the release of the substrate into the cytoplasm (inward-facing occluded) (Fig. 5). The isomerization between these states involves a combination of a cross-membrane movement and rotation of the transport domain, comprised of the substrate-binding transporter core and peripheral TM3 and TM6. During this movement, the lipid-facing hydrophobic transmembrane segments, TM3 and TM6, traverse the bilayer directly, moving towards the cytoplasm by ~12 Å. In contrast, passage of the relatively polar HP1 and HP2, with their tips crossing as much as 20 Å of the lipid bilayer, is facilitated by the intra-protein track provided by the trimerization domain. The transport domains are expected to move stochastically and independently within each protomer, so that the crystal structure represents a statistically rare case, when all three transport domains are in the inward orientation. The rigid trimerization domain is anchored in the membrane and provides a counterbalance to the movements of the bulky transport domains, suggesting an explanation for the obligatory oligomeric assembly of SLC1 family of transporters.

Substrate binding/unbinding on either side of the membrane is associated with additional conformational changes, or gate openings. The crystallographic data and molecular dynamic simulations[9,37,38] suggest that substrate and ion dissociation on the extracellular side is associated with the opening of HP2, defining it as an extracellular gate. We hypothesize that in a functionally symmetrical manner, movements of HP1 allow for dissociation of the substrate and ions from the inward facing state, defining it as a bona fide intracellular gate. The alternate opening of the extracellular and intracellular gates

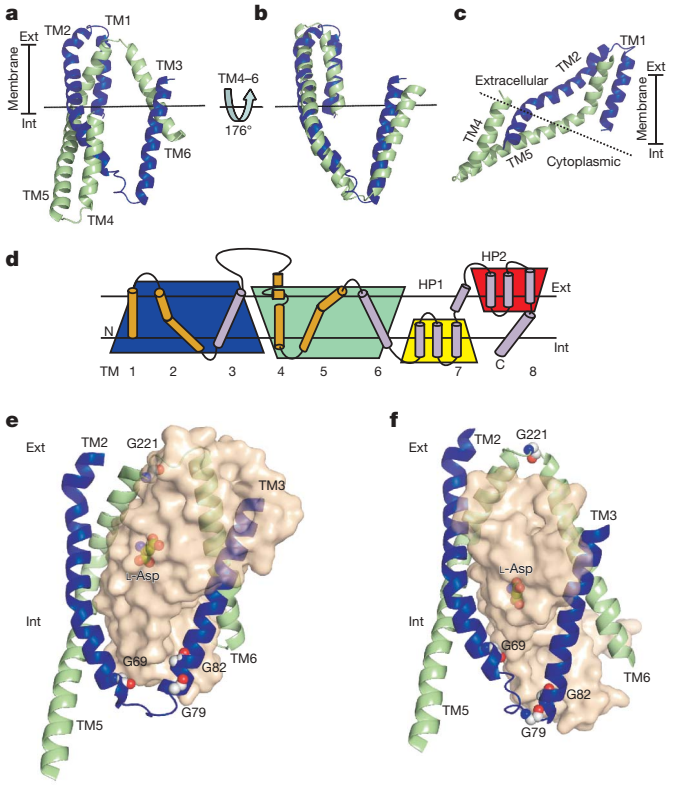



**Figure 4 | Amino-terminal inverted structural repeat.** **a**, **b**, Cartoon representation of TM1–TM3 (blue) and TM4–TM6 (green) in the $Glt_{Ph}$(55C/364$C_{Hg}$) viewed in the membrane plain (**a**), and their structural superposition (**b**). **c**, Symmetrical helices, TM1–TM2 and TM4–TM5, form the interaction surface within the transport domain, which is partitioned into intracellular and extracellular halves delineated by the dotted line. **d**, Schematic representation of $Glt_{Ph}$ trimerization (orange) and the transport (light blue) domains. Two inverted structural repeats are emphasized by blue and green and yellow and red trapezoids. **e**, **f**, Structure of TM2–TM3 and TM5–TM6 loops in wild-type $Glt_{Ph}$ (**e**) and $Glt_{Ph}$(55C/364$C_{Hg}$) (**f**). TM2–TM6 are shown in cartoon representation with TM4 omitted for clarity. The transporter core is shown in transparent surface representation. Bound L-Asp and conserved glycines are shown as spheres.

is strictly maintained: in the outward facing state, when HP2 opens to expose the substrate and ion binding sites to the extracellular solution, HP1 is secured in the closed state, packed against TM2 and TM5. Conversely, in the inward facing state HP2, the extracellular gate, is locked closed upon displacing HP1.

As has been suggested previously[39], coupling of the substrate transport to the energy of the ionic gradients is established via synergistic binding of substrate and sodium ions on the extracellular and intracellular sides of the membrane coupled to the closure of the corresponding gates. The steep sodium dependence of the substrate binding, observed both in wild-type $Glt_{Ph}$ (ref. 9) and in $Glt_{Ph}$(55C/364$C_{Hg}$) (Fig. 2d), is responsible for the differential affinity for the substrate on the extracellular side of the membrane, where sodium concentration is high, and the intracellular side, where it is low. We further propose that the isomerization between the outward and inward facing states is sodium independent and may simply be driven by the thermal energy alone. The observation that $Hg^{2+}$-mediated crosslinking of the K55C/A364C mutant is completed within seconds both in membranes and in detergent suggests that this transition is rapid and does not limit the rate of transport in $Glt_{Ph}$, which has a turnover time of ~3 min at ambient temperatures[40]. Finally, we hypothesize that a similar isomerization reaction may also occur in the *apo*-transporter to complete the transport cycle. The structural re-arrangements within the transport domain, which would allow for the closure of the extracellular and the intracellular gates in the absence of bound substrate and ions, remain to be elucidated.

*Note added in proof*: During the review of this paper, we became aware of a modelling study of the inward facing state based on the symmetry considerations[51]. It recapitulates several features of the crystal structure, and in particular proposes that HP2 packs against TM2 and TM5.

## METHODS SUMMARY

**Cysteine crosslinking.** $Glt_{Ph}$(K55C/C321A/A364C) was expressed as a $His_8$ fusion protein and purified as described previously[8]. Transporter samples were exchanged by size-exclusion chromatography into buffer, containing 10 mM HEPES/NaOH or KOH, pH 7.4, 1 mM *n*-dodecyl-β-D-maltopyranoside and either 100 mM NaCl and 100 μM L-Asp or 100 mM KCl. Crosslinking was initiated by addition of 1:2 molar ratio of $Cu^{2+}$ and 1,10-phenanthroline or $HgCl_2$ at indicated concentrations. Reactions were quenched with 100 mM *N*-ethyl maleimide before SDS–polyacrylamide gel electrophoresis analysis. Crude *E. coli* membranes were isolated by centrifugation, washed either in a $Na^+$/ L-Asp-containing or free buffer and crosslinked as in detergent. Protein bands were visualized by western blotting using antibodies against histidine tag.
**Crystallography.** The K55C/C321A/A364C mutation was introduced within a heptahistidine mutant of $Glt_{Ph}$, used in earlier crystallographic studies[8,9], to which we refer as 'wild type' throughout the text for brevity. Purified protein was crosslinked in the presence of 10-fold molar excess of $HgCl_2$, dialysed against buffer containing 10 mM HEPES/NaOH, 7 mM *n*-decyl-β-D-maltopyranoside, 100 mM NaCl and 100 μM L-Asp, diluted to the final concentration of 2–4 mg $ml^{-1}$ and supplemented with 0.5 mM *E. coli* total polar lipid extract and 100 mM NaBr. Protein solution was mixed at 1:1 (vol.:vol.) ratio with the reservoir solution, containing 100 mM MES, pH 5.0, 18–20% PEG 350 MME and 200 mM $CaCl_2$, and crystallized at 4 °C by hanging-drop vapour diffusion. Crystals were cryoprotected by allowing the drop to dry until its volume was reduced by 50%. SeMet substituted protein was expressed as described previously[8] and crystallized as above. Diffraction data were indexed, integrated and scaled using the HKL-2000 package[41]. Further analyses were performed using

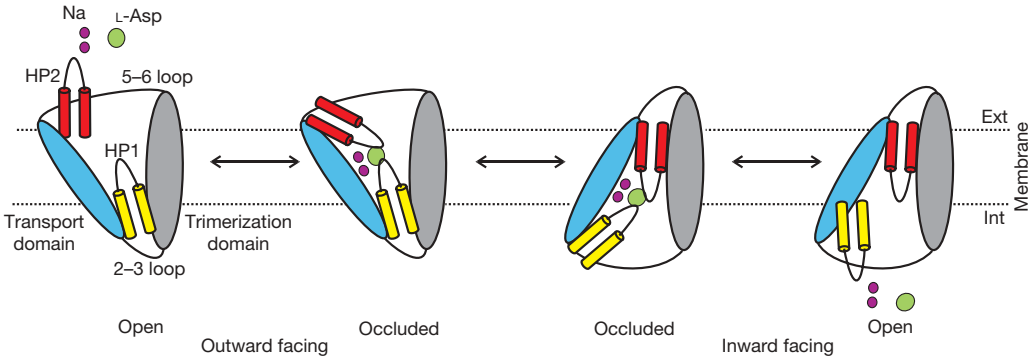



**Figure 5 | Schematic transport mechanism.** Shown is a single transporter protomer. Substrate and sodium binding to the outward and inward facing states is coupled to the closure of the extracellular and intracellular gates, HP2 (red) and HP1 (yellow), respectively. Isomerization between the outward and inward facing occluded states occurs upon movement of the transport domain (blue), relative to the trimerization domain (grey). The inward facing open state has not been structurally characterized and is hypothetical.

CCP4 programs[42]. Initial phases were obtained using Phaser[43], and the protein model built manually in Coot[44] and refined using REFMAC[42] with TLS[45] and three-fold NCS restraints.

**Received 24 July; accepted 29 October 2009.**
**Published online 18 November 2009.**

1. Danbolt, N. C. Glutamate uptake. *Prog. Neurobiol.* **65,** 1–105 (2001).
2. Saier, M. H. Jr, Tran, C. V. & Barabote, R. D. TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.* **34,** D181–D186 (2006).
3. Mitchell, P. A general theory of membrane transport from studies of bacteria. *Nature* **180,** 134–136 (1957).
4. Patlak, C. S. Contributions to the theory of active transport: II. The gate type non-carrier mechanism and generalizations concerning tracer flow, efficiency, and measurement of energy expenditure. *Bull. Math. Biol.* **19,** 209–235 (1957).
5. Jardetzky, O. Simple allosteric model for membrane pumps. *Nature* **211,** 969–970 (1966).
6. Sobczak, I. & Lolkema, J. S. Structural and mechanistic diversity of secondary transporters. *Curr. Opin. Microbiol.* **8,** 161–167 (2005).
7. Krishnamurthy, H., Piscitelli, C. L. & Gouaux, E. Unlocking the molecular secrets of sodium-coupled transporters. *Nature* **459,** 347–355 (2009).
8. Yernool, D., Boudker, O., Jin, Y. & Gouaux, E. Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. *Nature* **431,** 811–818 (2004).
9. Boudker, O., Ryan, R. M., Yernool, D., Shimamoto, K. & Gouaux, E. Coupling substrate and ion binding to extracellular gate of a sodium-dependent aspartate transporter. *Nature* **445,** 387–393 (2007).
10. Yernool, D., Boudker, O., Folta-Stogniew, E. & Gouaux, E. Trimeric subunit stoichiometry of the glutamate transporters from *Bacillus caldotenax* and *Bacillus stearothermophilus*. *Biochemistry* **42,** 12981–12988 (2003).
11. Gendreau, S. *et al.* A trimeric quaternary structure is conserved in bacterial and human glutamate transporters. *J. Biol. Chem.* **279,** 39505–39512 (2004).
12. Raunser, S. *et al.* Structure and function of prokaryotic glutamate transporters from *Escherichia coli* and *Pyrococcus horikoshii*. *Biochemistry* **45,** 12796–12805 (2006).
13. Koch, H. P. & Larsson, H. P. Small-scale molecular motions accomplish glutamate uptake in human glutamate transporters. *J. Neurosci.* **25,** 1730–1736 (2005).
14. Grewer, C. *et al.* Individual subunits of the glutamate transporter EAAC1 homotrimer function independently of each other. *Biochemistry* **44,** 11913–11923 (2005).
15. Leary, G. P., Stone, E. F., Holley, D. C. & Kavanaugh, M. P. The glutamate and chloride permeation pathways are colocalized in individual neuronal glutamate transporter subunits. *J. Neurosci.* **27,** 2938–2942 (2007).
16. Koch, H. P., Brown, R. L. & Larsson, H. P. The glutamate-activated anion conductance in excitatory amino acid transporters is gated independently by the individual subunits. *J. Neurosci.* **27,** 2943–2947 (2007).
17. Groeneveld, M. & Slotboom, D. J. Rigidity of the subunit interfaces of the trimeric glutamate transporter GltT during translocation. *J. Mol. Biol.* **372,** 565–570 (2007).
18. Shimamoto, K. *et al.* DL-threo-β-benzyloxyaspartate, a potent blocker of excitatory amino acid transporters. *Mol. Pharmacol.* **53,** 195–201 (1998).
19. Ryan, R. M., Mitrovic, A. D. & Vandenberg, R. J. The chloride permeation pathway of a glutamate transporter and its proximity to the glutamate translocation pathway. *J. Biol. Chem.* **279,** 20742–20751 (2004).
20. Manceau, A. & Nagy, K. L. Relationships between Hg(II)-S bond distance and Hg(II) coordination in thiolates. *Dalton Trans.* 1421–1425 (2008).
21. Dieckmann, C. R. *et al. De novo* design of mercury-binding two- and three-helical bundles. *J. Am. Chem. Soc.* **119,** 6195–6196 (1997).
22. Rosenzweig, A. C. *et al.* Crystal structure of the Atx1 metallochaperone protein at 1.02 Å resolution. *Structure* **7,** 605–617 (1999).
23. Sadeghi, M., Naderi-Manesh, H., Zarrabi, M. & Ranjbar, B. Effective factors in thermostability of thermophilic proteins. *Biophys. Chem.* **119,** 256–270 (2006).
24. Grunewald, M., Menaker, D. & Kanner, B. I. Cysteine-scanning mutagenesis reveals a conformationally sensitive reentrant pore-loop in the glutamate transporter GLT-1. *J. Biol. Chem.* **277,** 26074–26080 (2002).
25. Borre, L., Kavanaugh, M. P. & Kanner, B. I. Dynamic equilibrium between coupled and uncoupled modes of a neuronal glutamate transporter. *J. Biol. Chem.* **277,** 13501–13507 (2002).
26. Seal, R. P., Shigeri, Y., Eliasof, S., Leighton, B. H. & Amara, S. G. Sulfhydryl modification of V449C in the glutamate transporter EAAT1 abolishes substrate transport but not the substrate-gated anion conductance. *Proc. Natl Acad. Sci. USA* **98,** 15324–15329 (2001).
27. Ryan, R. M. & Vandenberg, R. J. Distinct conformational states mediate the transport and anion channel properties of the glutamate transporter EAAT-1. *J. Biol. Chem.* **277,** 13494–13500 (2002).
28. Leighton, B. H., Seal, R. P., Shimamoto, K. & Amara, S. G. A hydrophobic domain in glutamate transporters forms an extracellular helix associated with the permeation pathway for substrates. *J. Biol. Chem.* **277,** 29847–29855 (2002).
29. Slotboom, D. J., Sobczak, I., Konings, W. N. & Lolkema, J. S. A conserved serine-rich stretch in the glutamate transporter family forms a substrate-sensitive reentrant loop. *Proc. Natl Acad. Sci. USA* **96,** 14282–14287 (1999).
30. Grunewald, M., Bendahan, A. & Kanner, B. I. Biotinylation of single cysteine mutants of the glutamate transporter GLT-1 from rat brain reveals its unusual topology. *Neuron* **21,** 623–632 (1998).
31. Seal, R. P., Leighton, B. H. & Amara, S. G. Transmembrane topology mapping using biotin-containing sulfhydryl reagents. *Methods Enzymol.* **296,** 318–331 (1998).
32. Shlaifer, I. & Kanner, B. I. Conformationally sensitive reactivity to permeant sulfhydryl reagents of cysteine residues engineered into helical hairpin 1 of the glutamate transporter GLT-1. *Mol. Pharmacol.* **71,** 1341–1348 (2007).
33. Abramson, J. *et al.* Structure and mechanism of the lactose permease of *Escherichia coli*. *Science* **301,** 610–615 (2003).
34. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of $Na^+/Cl^-$-dependent neurotransmitter transporters. *Nature* **437,** 215–223 (2005).
35. Hunte, C. *et al.* Structure of a $Na^+/H^+$ antiporter and insights into mechanism of action and regulation by pH. *Nature* **435,** 1197–1202 (2005).
36. Shachnai, L., Shimamoto, K. & Kanner, B. I. Sulfhydryl modification of cysteine mutants of a neuronal glutamate transporter reveals an inverse relationship between sodium dependent conformational changes and the glutamate-gated anion conductance. *Neuropharmacology* **49,** 862–871 (2005).
37. Huang, Z. & Tajkhorshid, E. Dynamics of the extracellular gate and ion-substrate coupling in the glutamate transporter. *Biophys. J.* **95,** 2292–2300 (2008).
38. Shrivastava, I. H., Jiang, J., Amara, S. G. & Bahar, I. Time-resolved mechanism of extracellular gate opening and substrate binding in glutamate transporter. *J. Biol. Chem.* **283,** 28680–28690 (2008).
39. Gouaux, E. The molecular logic of sodium-coupled neurotransmitter transporters. *Phil. Trans. R. Soc. B* **364,** 149–154 (2009).
40. Ryan, R. M., Compton, E. L. & Mindell, J. A. Functional characterization of a $Na^+$-dependent aspartate transporter from *Pyrococcus horikoshii*. *J. Biol. Chem.* **284,** 17540–17548 (2009).
41. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 308–326 (1997).
42. Collaborative Computational Project, number 4. The CCP4 Suite: Programs for X-ray crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
43. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
44. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
45. Winn, M. D., Isupov, M. N. & Murshudov, G. N. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D* **57,** 122–133 (2001).
46. DeLano, W. L. *The PyMOL Molecular Graphics System* (DeLano Scientific LLC, 2008).
47. Saha, R. P., Bahadur, R. P., Pal, A., Mandal, S. & Chakrabarti, P. ProFace: a server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Struct. Biol.* **6,** 11 (2006).
48. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33,** W299–W302 (2005).
49. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36,** D281–D288 (2008).
50. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
51. Crisman, T. J., Qu, S., Kanner, B. I. & Forrest, L. R. The inward-facing conformation of glutamate transporters as revealed by their inverted-topology structural repeats. *Proc. Natl. Acad. Sci. USA* (in the press).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank D. Patel and H. Li for the help with isothermal titration calorimetry and H. Weinstein for constructive criticism. X-ray diffraction data were measured at X25 beamline at the National Synchrotron Light Source. This work was supported by the National Institute of Health (O.B.) and by a Jane Coffin Childs Memorial Fund postdoctoral fellowship (N.R.).

**Author Contributions** N.R. and O.B. designed the project. N.R. performed protein purification, crosslinking, crystallization and structure determination. C.G. performed cloning, cell culture and protein purification. N.R. and O.B. wrote the manuscript.

**Author Information** The coordinates for the structure and the structure factors are deposited in the Protein Data Bank under accession code 3KBC. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to O.B. (olb2003@med.cornell.edu).

Vol 462|17 December 2009|doi:10.1038/nature08593

ARTICLES

# The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress

Joanna R. Morris[1]*, Chris Boutell[2]*, Melanie Keppler[3]*, Ruth Densham[1], Daniel Weekes[1], Amin Alamshah[1], Laura Butler[1], Yaron Galanty[4], Laurent Pangon[1], Tai Kiuchi[3], Tony Ng[3] & Ellen Solomon[1]

**Mutations in *BRCA1* are associated with a high risk of breast and ovarian cancer. BRCA1 participates in the DNA damage response and acts as a ubiquitin ligase. However, its regulation remains poorly understood. Here we report that BRCA1 is modified by small ubiquitin-like modifier (SUMO) in response to genotoxic stress, and co-localizes at sites of DNA damage with SUMO1, SUMO2/3 and the SUMO-conjugating enzyme Ubc9. PIAS SUMO E3 ligases co-localize with and modulate SUMO modification of BRCA1, and are required for BRCA1 ubiquitin ligase activity in cells. *In vitro* SUMO modification of the BRCA1/BARD1 heterodimer greatly increases its ligase activity, identifying it as a SUMO-regulated ubiquitin ligase (SRUbL). Further, PIAS SUMO ligases are required for complete accumulation of double-stranded DNA (dsDNA) damage-repair proteins subsequent to RNF8 accrual, and for proficient double-strand break repair. These data demonstrate that the SUMOylation pathway plays a significant role in mammalian DNA damage response.**

The amino (N) terminus of BRCA1 has a RING domain that interacts with ubiquitin-conjugating enzymes and is required for its ubiquitin ligase activity[1–3]. Many disease-causing mutations are found within this region, and loss of the ligase activity is associated with susceptibility to breast cancer[4]. We have previously shown that BRCA1-dependent ubiquitin conjugates are generated at sites of DNA damage repair in human cells[5]. Although the substrate(s) of the BRCA1 ubiquitin ligase activity remains controversial[6], and its role at this location unclear[7], the ligase activity itself is highly conserved and damage-associated ubiquitin conjugates are also formed by *Caenorhabditis elegans*[8] and *Gallus gallus*[9] homologues of BRCA1.

BRCA1 recruitment to chromatin at sites of DNA damage occurs through a complex cascade of protein modifications and interactions, and BRCA1 is the third in a sequence of ubiquitin ligases recruited to such sites[10]. Binding of the mediator of DNA damage checkpoint 1 (MDC1) protein to the phosphorylated tail of histone H2AX (γ-H2AX) at sites of DNA breakage recruits the ubiquitin ligase RNF8[11–13], which generates ubiquitin chains bound by RAP80:ABRA1, which in turn recruits BRCA1 through its carboxy (C) terminus[14–20]. The activity of the second ubiquitin ligase, RNF168, maintains the ubiquitin chain signal initiated by RNF8 and thus helps retain BRCA1 at these sites[10,21].

SUMOylation of substrates is catalysed by a cascade of enzymes: the activities of the E1 SUMO-activating enzyme (SAE1/SAE2), the E2-conjugating enzyme (Ubc9) and E3 SUMO ligases result in an isopeptide bond between the target lysine and the activated SUMO carboxyl terminus (reviewed in ref. 22). In vertebrates, three SUMO isoforms, SUMO1, SUMO2 and SUMO3, are expressed. SUMO2 and SUMO3 differ by three N-terminal residues, and form a distinct subfamily known as SUMO2/3. Ubc9 and SUMO E3 enzymes have been implicated in the DNA damage response in human cells and animal models[23–27]. In *C. elegans*, ce-ubc9 interacts with ce-Bard1, the N-terminal binding partner of ce-Brca1[28]. The interaction of BRCA1 with free, non-conjugated SUMO1 has been shown to decrease its transcriptional activity[29]. Here we investigated the previously unaddressed role of the SUMO pathway in the regulation of BRCA1 and in the DNA damage response.

## SUMO conjugation in response to damage

We noted that after treatments with genotoxic agents (irradiation, cisplatin and hydroxyurea), SUMO isoforms, SUMO1 and SUMO2/3 localized to sub-nuclear damage repair foci marked with γ-H2AX and BRCA1 (Fig. 1a–c, Supplementary Fig. 2a and data not shown). We also saw increased fluorescence resonance energy transfer (FRET) between green fluorescent protein (GFP)–BRCA1 and red fluorescent protein (RFP)–SUMO1, measured by fluorescence lifetime imaging microscopy (FLIM), after treatment of cells with the genotoxic agents cisplatin, irradiation, hydroxyurea and epirubicin (Supplementary Fig. 2b, c), indicating increased protein–protein interaction[30–34]. Increased FRET populations were observed in regions largely coincident with γ-H2AX foci in hydroxyurea-treated cells, indicating BRCA1–SUMO1 interaction at or close to γ-H2AX-decorated chromatin (Fig. 1d). The interaction was dependent on conjugation, as a SUMO mutant with a C-terminal di-glycine substitution (GG to GA) exhibited lower FRET with GFP–BRCA1 (Fig. 1e, f). Consistent with this, immunoprecipitation of endogenous BRCA1 from hydroxyurea-treated cells co-purified high molecular mass endogenous SUMO conjugates (Supplementary Fig. 2d), suggesting either SUMOylation of BRCA1 itself or interaction of BRCA1 with large, SUMO-modified proteins. *De novo* SUMOylation at sites of DNA damage would require localization of the SUMO-conjugating enzyme, Ubc9, to these sites. We found that Ubc9–GFP co-localized both with

[1]Department of Medical and Molecular Genetics, King's College London, Guy's Medical School Campus, London SE1 9RT, UK. [2]MRC Virology Unit, Church Street, Glasgow G11 5JR, Scotland, UK. [3]Richard Dimbleby Department of Cancer Research, Randall Division of Cell and Molecular Biophysics, King's College London, Guy's Medical School Campus, London SE1 1UL, UK. [4]The Wellcome Trust/Cancer Research UK Gurdon Institute and Department of Zoology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK.
*These authors contributed equally to this work.

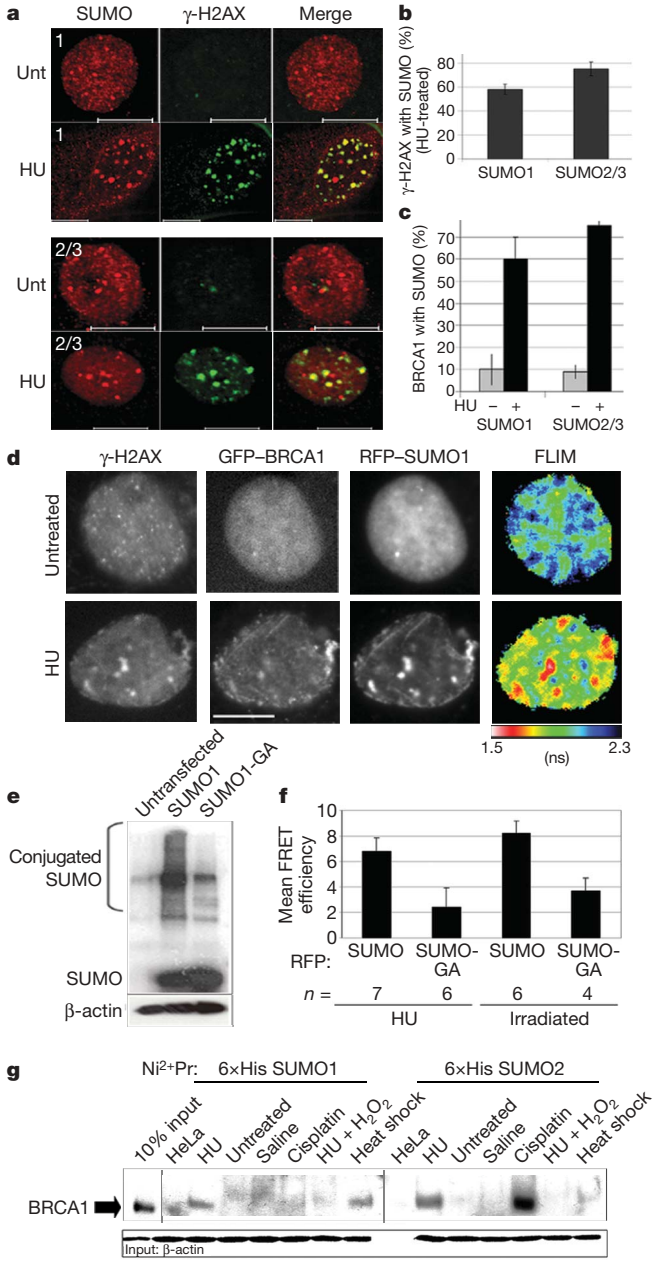



**Figure 1 | The SUMO-conjugation machinery locates to sites of DNA damage, and BRCA1 is modified by SUMO after genotoxic stress.** COS-7 cells with (HU) or without (Unt) hydroxyurea treatment stained with anti-SUMO isoforms and γ-H2AX (**a**) or BRCA1 and counted in **b** and **c**. Error bars, s.d.; $n > 30$ cells per condition. Scale bars, 10 μm throughout the figures. **d**, BRCA1–SUMO1 interaction at sites of genotoxic stress. Images of γ-H2AX, GFP and RFP multiphoton intensity and FLIM in transfected cells with or without hydroxyurea. FRET shortens the GFP-fluorescence lifetime in orange to red. **e**, Myc-SUMO1-GA conjugates poorly compared with wild-type myc-SUMO1 in denaturing SDS–polyacrylamide gel electrophoresis immunoblotted with anti-myc. **f**, FRET efficiency of RFP-SUMO1-GA or RFP-SUMO1 with GFP–BRCA1 in treated cells. Bars, s.e.m. **g**, SUMOylation of BRCA1. Nickel precipitation (Pr, $Ni^{2+}$) from untransfected (HeLa) and 6×His-SUMO-expressing cells treated with the agents shown.

γ-H2AX and BRCA1 in treated cells and showed increased FRET with RFP–BRCA1 (Supplementary Fig. 2e, f).

These data are consistent with modification of BRCA1 and/or closely associated proteins. As SUMO proteins are covalently linked to target proteins, we purified hexahistidine (6×His)-tagged SUMO conjugates using nickel-charged beads in highly denaturing conditions, and probed for BRCA1. BRCA1 was enriched on nickel after cell treatment with hydroxyurea, cisplatin and heat shock but was absent from untreated or untransfected cells, or cells treated with hydroxyurea and then low concentration $H_2O_2$ (Fig. 1g). ($H_2O_2$ cross-links Ubc9 and SAE2, preventing SUMO conjugation in HeLa cells[35], indicating that like most known SUMO targets, the modification of BRCA1 is rapidly processed.) Both SUMO isoforms conjugate to BRCA1, but more BRCA1 was purified with SUMO2 than SUMO1, particularly after cisplatin treatment. The effect of genotoxic agents on total cellular SUMO conjugates is shown in Supplementary Fig. 2g.

Together, these data indicate that the SUMO conjugation pathway forms part of the mammalian response to DNA damage, as Ubc9 and the SUMO proteins interact with at least one DNA damage-regulated SUMOylation target, BRCA1, at sites of genotoxic stress labelled by γ-H2AX.

## PIAS SUMO ligases in the damage response

Members of the protein inhibitor of activated signal transducer and activator of transcription (PIAS) family of SUMO E3 ligases and the Mms21(NSE2) SUMO ligase are found in foci within the nucleus[36], and have been reported to play a role in the DNA damage response[24–27]. Short interfering RNA (siRNA) depletion of PIAS1 and PIAS4 impaired 6×His SUMO1 and SUMO2 modification of endogenous BRCA1 in hydroxyurea-treated cells, unlike depletion of MMS21, PIAS2 or PIAS3 (Fig. 2a and data not shown). In untreated cells, depletion of PIAS1 resulted in increased SUMO2-conjugated BRCA1: this was dependent on PIAS4 because depletion of PIAS1 and PIAS4 inhibited the modification (Fig. 2a). Quantitative PCR with reverse transcription (RT–PCR) showed that siRNAs to each SUMO ligase were E3 specific, and ligase depletion had no impact on steady-state BRCA1 protein levels (Supplementary Fig. 3a, b). Ectopic expression of PIAS1 and PIAS4 (but not MMS21) increased BRCA1–SUMO interaction, as measured by FLIM (Fig. 2b) and 6×His-SUMO-conjugated BRCA1 (Fig. 2c). Increased expression or depletion of PIAS ligases had no obvious impact on total SUMO1 conjugates, but SUMO2 conjugates (in 6×His-SUMO2 cells) were decreased on PIAS1 depletion, and increased after ectopic expression of PIAS1 and PIAS4 (Fig. 2a and

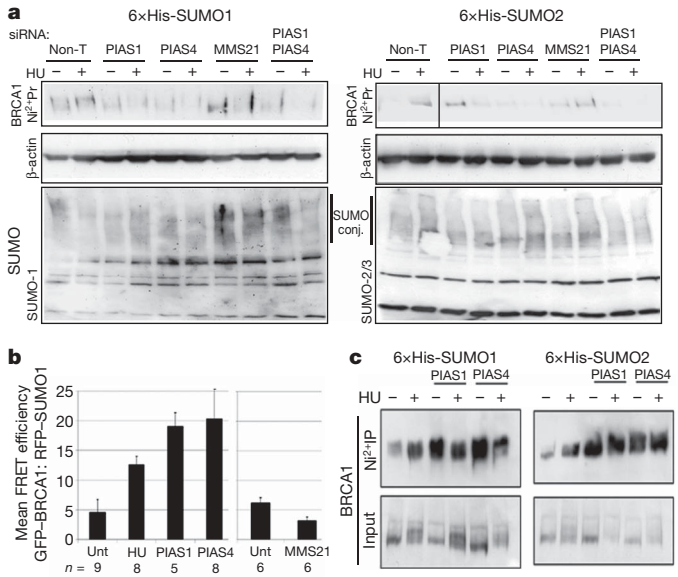



**Figure 2 | PIAS E3 SUMO ligases modulate BRCA1 SUMOylation.** **a**, PIAS ligases are required for BRCA1 SUMOylation. Nickel elutant from untreated or hydroxyurea-treated 6×His-SUMO-expressing cells transfected with siRNA to transcripts shown (Non-T, non-targeting) and probed with antibodies shown. SUMO conj, SUMO conjugates. **b**, PIAS ligases increase BRCA1–SUMO1 interaction. FRET efficiency of RFP-SUMO1 with GFP–BRCA1 in untreated (Unt) cells or transfected with SUMO ligases shown, or cells treated with hydroxyurea. Bars, s.e.m. **c**, PIAS ligases increase BRCA1 SUMOylation. Nickel elutant from 6×His-SUMO-expressing cells transfected with PIAS expression constructs.

Supplementary Fig. 3c). Thus BRCA1 is one of many substrates modified by SUMO2 in response to PIAS activity but is one of a smaller population of SUMO1 substrates regulated by PIAS proteins.

We assessed the impact of SUMO-ligase depletion on BRCA1 localization in cells and found that loss of PIAS1 and PIAS4, but not MMS21, PIAS2 or PIAS3 SUMO ligases, reduced its ability to localize to γ-H2AX in hydroxyurea-treated cells (Fig. 3a, b and data not shown). The introduction of siRNA-resistant PIAS proteins at low expression levels restored the ability of endogenous BRCA1 to co-localize with γ-H2AX (Supplementary Fig. 4a). PIAS1, PIAS4 and MMS21 ligases expressed in cells co-localized with BRCA1. High expression of PIAS1 or PIAS4, but not MMS21, together with BRCA1 and SUMO1 or SUMO2, resulted in exaggerated GFP–BRCA1 foci (Supplementary Fig. 4c). The formation of exaggerated foci was dependent on the SUMO ligase activity, because an inactive PIAS1 RING mutant[37] inhibited the ability of PIAS1 both to induce increased BRCA1–SUMO1 interaction in cells (Supplementary Fig. 4b) and to cause exaggerated BRCA1 foci (Supplementary Fig. 4c, d). Together, these data indicate that PIAS1 and PIAS4 SUMO ligases modulate and are required for normal accumulation of BRCA1 at sites of genotoxic stress.

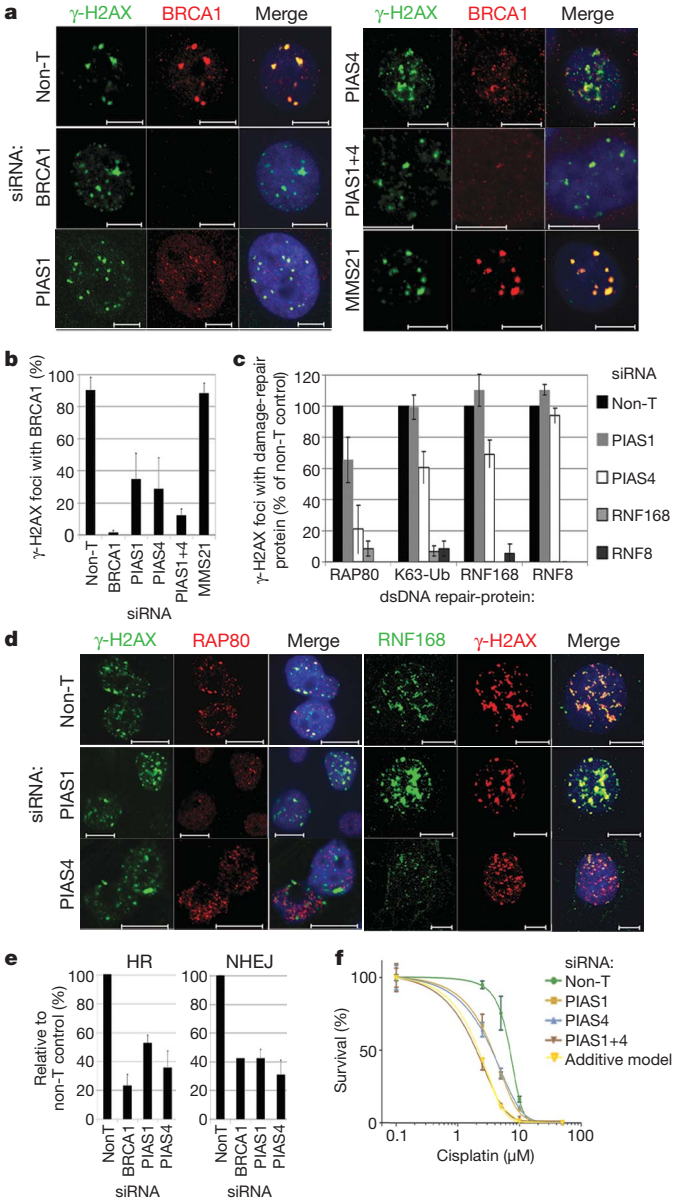



**Figure 3 | PIAS E3 SUMO ligases influence accumulation of DNA damage-repair protein and are required for dsDNA break repair. a**, BRCA1 accumulation to γ-H2AX in hydroxyurea-treated cells transfected with siRNAs indicated and counted in **b** (bars, s.d.; $n > 30$ cells per condition). **c**, Hydroxyurea and siRNA treated cells scored for γ-H2AX foci with accumulated dsDNA damage proteins (bars, s.d.; $n > 30$ cells per condition) and representative immunofluorescence images shown in **d**. **e**, Homologous recombination (HR) and non-homologous end joining (NHEJ) assayed in siRNA transfected cells bearing integrated gene conversion and end-joining substrates after I-sceI-induced dsDNA break. Bars, s.d. **f**, Colony survival of siRNA-transfected cells exposed to cisplatin. Combined sensitivity of PIAS1 + PIAS4 was characterized as additive by the Bliss independence model[50].

To examine whether impaired BRCA1 accumulation following PIAS depletion is likely to be direct or indirect, we investigated the integrity of the upstream accumulation cascade by examining RAP80, K63-linked ubiquitin (K63-Ub), RNF168 and RNF8, all proteins necessary for BRCA1 recruitment to sites of genotoxic stress[10–14,16,18–21]. PIAS4 was required for normal accumulation of proteins subsequent to RNF8, affecting RNF168, K63-Ub, RAP80 and BRCA1, whereas PIAS1 was required for complete accumulation of proteins subsequent to the generation of K63-linked ubiquitin, affecting RAP80 and BRCA1 (Fig. 3c, d and Supplementary Fig. 4e). Thus, although BRCA1 is SUMO modified in a PIAS-dependent manner, these data indicate that its accumulation is regulated through an indirect mechanism involving earlier-arriving proteins.

Consistent with the requirement on PIAS proteins for BRCA1 accumulation, depletion of these ligases, like depletion of BRCA1, reduced homologous recombination and non-homologous end-joining repair of double-strand breaks, and increased cellular sensitivity to cisplatin (Fig. 3e, f), indicating that these SUMO ligases are required for the full response to DNA damage.

## SUMO conjugation influences BRCA1 ligase

In cells, ubiquitin conjugates detected by the monoclonal antibody FK2 are lost from sites of DNA damage after depletion of many proteins involved in accumulation of dsDNA repair protein, including the ubiquitin ligases RNF8, RNF168 and BRCA1 (refs 5, 8–10, 21, 38). FK2-ubiquitin accumulation to γ-H2AX was also reduced after depletion of PIAS1 and PIAS4 (data not shown). Because PIAS1 depletion does not impair RNF168/K63-Ub accumulation (Fig. 3c), this suggests that the impact on FK2-ubquitin is independent or subsequent to RNF8/RNF168 activity. FK2-ubiquitin accumulation was also reduced in the small population of PIAS-depleted cells that retained some BRCA1 foci (Fig. 4a, b), suggesting disruption of an activity subsequent to BRCA1 accumulation.

Ectopic expression of BRCA1/BARD1 in cells is able to increase FK2-ubiquitin conjugate staining above levels detected in surrounding, non-transfected cells, in a manner dependent on a functional RING domain[5]. This ability was absent in PIAS1/4-depleted cells (Supplementary Fig. 5a), indicating that the BRCA1 ubiquitin ligase activity is reduced. Similarly the co-localization of K6-linked ubiquitin (catalysed by BRCA1 (refs 5, 39, 40)) with γ-H2AX was impaired in BRCA1-, PIAS1- or PIAS4-depleted cells, consistent with loss of BRCA1 ligase activity (Supplementary Fig. 5b).

SUMO conjugation frequently, but not always, occurs on lysines in the consensus motif, 'ΨKxE', where Ψ is an aliphatic residue. The two highest-scoring motifs in BRCA1 are residues K109 and K119, adjacent to the RING domain (Fig. 4c). The interaction of BRCA1 K109→R with RFP–SUMO1 was comparable to that of wild-type BRCA1 (data not shown). However, substitutions of the K119 motif (K119→R or E121→A) reduced interaction with RFP–SUMO1 (Fig. 4d). These mutations, like the RING mutation, C61G, reduced the ability of ectopic BRCA1 to induce increased levels of co-localizing FK2 ubiquitin conjugates in cells (Fig. 4e, f), consistent with the effects of PIAS1 and PIAS4 ligase depletion.

These observations imply that the SUMOylation pathway acts directly on the BRCA1/BARD1 ligase. To test this, we explored BRCA1 SUMO modification *in vitro* (Supplementary Fig. 5c, d). Titration of unmodified against SUMO1-modified heterodimer showed that the modified form had increased ubiquitin ligase activity, generating 10–20 times more conjugated ubiquitin (Fig. 4g).

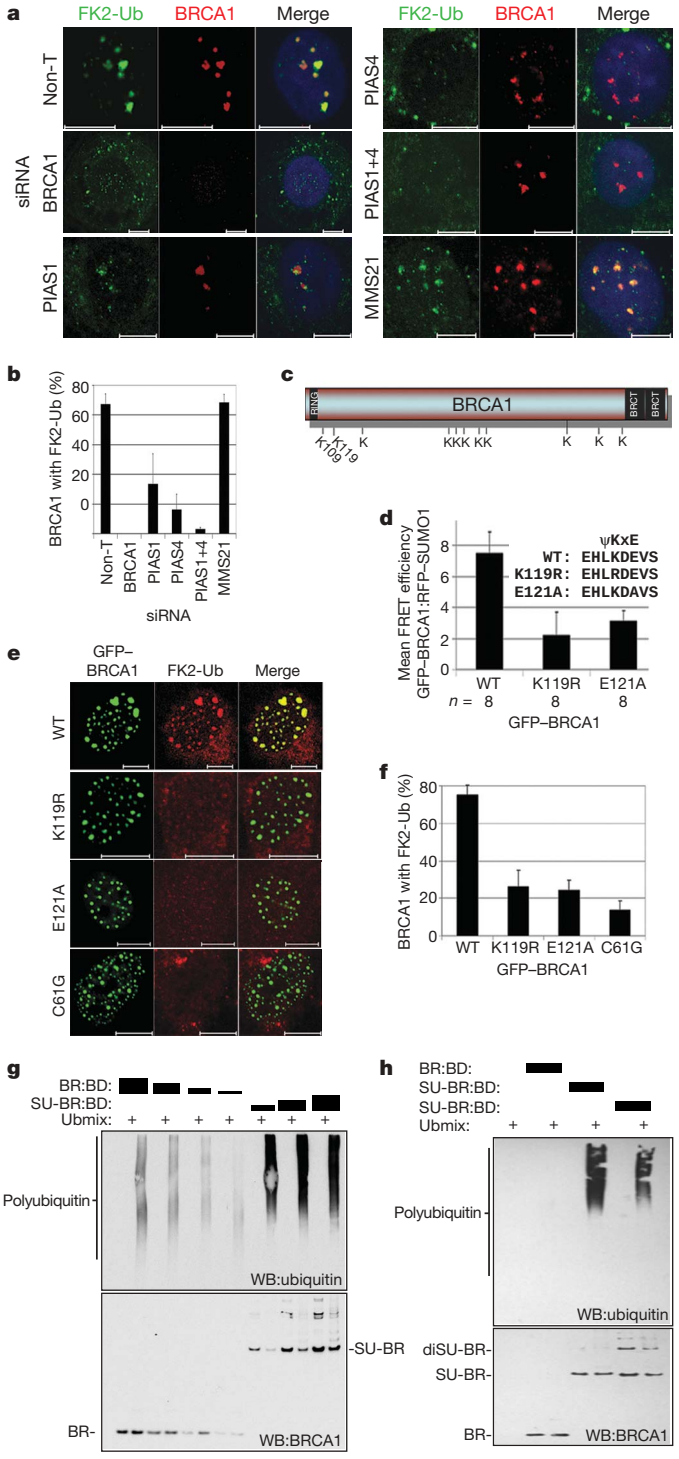



**Figure 4 | The SUMO pathway regulates BRCA1 ubiquitin ligase activity.** **a**, FK2-ubiquitin accumulation with BRCA1 in siRNA and hydroxyurea-treated cells and counted in **b** (bars, s.d.; $n > 30$ cells per condition). **c**, Illustration of BRCA1 motifs and consensus SUMO sites (K) identified by both abgent.com/doc/sumoplot and bioinformatics.lcd-ustc.org/sumosp programs; K119 and K109 rank highest in both. **d**, BRCA1 K119 consensus is required for interaction with SUMO1 in hydroxyurea-treated cells. Bars, s.e.m. **e**, Cells expressing SUMO consensus or RING-C61G mutants stained with low titration FK2 and scored in **f** (bars, s.d.; $n > 30$ cells per condition). WT, wild type. **g**, **h**, SUMO-modification increases the ubiquitin ligase activity of $BRCA1_{1–147}$–$BARD1_{26–142}$. BR, BRCA1; BD, BARD1; SU1, SUMO1; SU2, SUMO2. Ubmix indicates ubiquitin conjugation components except the heterodimer.

Comparison of unmodified heterodimer with SUMO1 and SUMO2 modified forms, at a concentration where the unmodified proteins had little detectable activity, showed that the SUMO modification enhanced ubiquitin ligase activity independent of isoform specificity (Fig. 4h).

## Discussion

SUMO modification increases BRCA1 ubiquitin ligase activity *in vitro*, consistent with the requirement in cells for PIAS SUMO E3 ligases and for an N-terminal SUMO modification consensus site, thus identifying BRCA1 as a SUMO-regulated ubiquitin ligase (SRUbL). SUMO modification of BRCA1 and occupation of the BRCA1 RING by ubiquitin E2 conjugating enzymes are concurrent, supported by the observation that the presence of several-fold molar excess of the ubiquitin E2, UbcH5a, has little impact on BRCA1 SUMO modification *in vitro* (data not shown). Similarly, mutations that inhibit BRCA1–ubiquitin E2 interactions (T77→M, I26→A, C61→G or the absence of BARD1 polypeptide) had no impact on BRCA1 SUMOylation, indicating that the BRCA1/BARD1 heterodimer RING domains are not required for SUMO pathway interaction (data not shown). Thus the simplest mechanism envisaged is that SUMO modification of BRCA1 increases the E3–E2 interface, through SUMO interaction with the E2 enzyme (possibly through SUMO interacting motifs). Based on *in vitro* investigations, other authors[41,42] have shown that auto-ubiquitylation of BRCA1 at positions C-terminal to its RING domain regulates ligase activity and E2 choice, although how BRCA1 SUMOylation and auto-ubiquitylation relate is yet to be clarified.

These data show that the PIAS SUMO ligases are necessary components of the mammalian response to double-strand breaks, required for homologous recombination and non-homologous end-joining, and that they influence BRCA1 accumulation through earlier-arriving proteins. However, the details of their regulation by the SUMO pathway remain to be determined. It is possible, for example, that like BRCA1, the other ubiquitin ligases in the pathway, RNF8 and RNF168, are also SRUbLs regulated by SUMO-modification (Supplementary Fig. 1).

The known post-translational modifications of BRCA1 now include phosphorylation, ubiquitylation and SUMOylation. Because the two features of BRCA1 activity regulated by the SUMO pathway, ubiquitin ligase activity and accumulation at sites of DNA damage, are also inhibited by some *BRCA1* mutations that predispose to breast and ovarian cancer[4,43], it seems highly likely that the SUMO pathway will be of relevance to cancer predisposition and development.

## METHODS SUMMARY

**Cell treatments.** Cells were treated with 10 μM cisplatin for 3 h followed by 16 h recovery, 16 h in 20 nM epirubicin, 3 mM hydroxyurea, or 3 mM hydroxyurea for 8 h followed by 1 mM $H_2O_2$ for 15 min, 15 min heat shock at 43 °C or exposure to 10 Gy irradiation using a caesium-137 source.

**SiRNA.** SMART Pool siRNAs (Dharmacon) used were: BRCA1 (L-003461-00), PIAS1 (L-008167-00), PIAS2 (L-009428-00), PIAS3 (L-004164-00) PIAS4 (L-006445-00) MMS21/NSE2 (L-018070-00), RNF168 (L007152-00-0005) and RNF8 (L-009600-00-0005), non-targeting control siRNA, confirmed to have minimal targeting of known genes (D-001810-10-05). The untranslated regions target sequences were as follows: GGCGAAGUUCACUGCGC (PIAS1), CAGAGGGAGGAGUGACC (PIAS4).

**Purification of 6×His tagged SUMO conjugates.** This was as described previously[44].

**Immunofluorescence microscopy.** This was performed as previously described[5].

**Antibodies.** The following antibodies were used: MS110 (Ab1, Calbiochem), FK-2 (Biomol), anti-Flag (M2) (Sigma), control rabbit IgG (Sigma), β-actin (Abcam), anti-c-myc, anti-SUMO1, anti-SUMO2/3 (Santa Cruz), γ-H2AX (Millipore and Abcam), anti-RNF8 (Abnova), anti-RNF168[10], anti-RAP80 (Bethyl) and anti-K63-ubiquitn (Millipore).

**Time-resolved multiphoton microscopy.** Measurements were undertaken with a modified system similar to that described previously[45]. Fluorescence lifetime imaging used time-correlated single-photon counting electronics (Becker & Hickl, SPC 830) collected through a bandpass filter centred at $\lambda = 510 \pm 10$ nm (Chroma). Excitation power was adjusted using a neutral density filter to photon counting rates $\sim 10^4$–$10^5$ photons $s^{-1}$, and acquisition times ~300 s at low excitation power were used. Imaging control and analysis used

custom software (CVI LabWindows)[46]. The Förster radius of the GFP and RFP pair used has been calculated as 4.7 nm (ref. 31).

**Repair sssays.** Homologous recombination and non-homologous end-joining cell assays were performed as previously descibed[47].

**Protein production and ubiquitin ligase assays.** These were as previously described[4].

***In vitro* SUMO conjugation assays.** These were as previously described[48,49].

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 17 May; accepted 19 October 2009.**


1. Lorick, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl Acad. Sci. USA* **96,** 11364–11369 (1999).
2. Hashizume, R. *et al.* The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. *J. Biol. Chem.* **276,** 14537–14540 (2001).
3. Brzovic, P. S. *et al.* Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. *Proc. Natl Acad. Sci. USA* **100,** 5646–5651 (2003).
4. Morris, J. R. *et al.* Genetic analysis of BRCA1 ubiquitin ligase activity and its relationship to breast cancer susceptibility. *Hum. Mol. Genet.* **15,** 599–606 (2006).
5. Morris, J. R. & Solomon, E. BRCA1: BARD1 induces the formation of conjugated ubiquitin structures, dependent on K6 of ubiquitin, in cells during DNA replication and repair. *Hum. Mol. Genet.* **13,** 807–817 (2004).
6. Morris, J. R. & Solomon, E. in *The Role of Genetics in Breast and Reproductive Cancers* (ed. Welcsh, P. L.) 75–92 (Springer Science+Business Media and Humana Press, 2009).
7. Reid, L. J. *et al.* E3 ligase activity of BRCA1 is not essential for mammalian cell viability or homology-directed repair of double-strand DNA breaks. *Proc. Natl Acad. Sci. USA* **105,** 20876–20881 (2008).
8. Polanowska, J., Martin, J. S., Garcia-Muse, T., Petalcorin, M. I. & Boulton, S. J. A conserved pathway to activate BRCA1-dependent ubiquitylation at DNA damage sites. *EMBO J.* **25,** 2178–2188 (2006).
9. Zhao, G. Y. *et al.* A critical role for the ubiquitin-conjugating enzyme Ubc13 in initiating homologous recombination. *Mol. Cell* **25,** 663–675 (2007).
10. Doil, C. *et al.* RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins. *Cell* **136,** 435–446 (2009).
11. Huen, M. S. *et al.* RNF8 transduces the DNA-damage signal via histone ubiquitylation and checkpoint protein assembly. *Cell* **131,** 901–914 (2007).
12. Kolas, N. K. *et al.* Orchestration of the DNA-damage response by the RNF8 ubiquitin ligase. *Science* **318,** 1637–1640 (2007).
13. Mailand, N. *et al.* RNF8 ubiquitylates histones at DNA double-strand breaks and promotes assembly of repair proteins. *Cell* **131,** 887–900 (2007).
14. Wang, B. & Elledge, S. J. Ubc13/Rnf8 ubiquitin ligases control foci formation of the Rap80/Abraxas/Brca1/Brcc36 complex in response to DNA damage. *Proc. Natl Acad. Sci. USA* **104,** 20759–20763 (2007).
15. Kim, H., Huang, J. & Chen, J. CCDC98 is a BRCA1-BRCT domain-binding protein involved in the DNA damage response. *Nature Struct. Mol. Biol.* **14,** 710–715 (2007).
16. Kim, H., Chen, J. & Yu, X. Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. *Science* **316,** 1202–1205 (2007).
17. Liu, Z., Wu, J. & Yu, X. CCDC98 targets BRCA1 to DNA damage sites. *Nature Struct. Mol. Biol.* **14,** 716–720 (2007).
18. Sobhian, B. *et al.* RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science* **316,** 1198–1202 (2007).
19. Wang, B. *et al.* Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. *Science* **316,** 1194–1198 (2007).
20. Yan, J. *et al.* The ubiquitin-interacting motif containing protein RAP80 interacts with BRCA1 and functions in DNA damage repair response. *Cancer Res.* **67,** 6647–6656 (2007).
21. Stewart, G. S. *et al.* The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. *Cell* **136,** 420–434 (2009).
22. Hay, R. T. SUMO: a history of modification. *Mol. Cell* **18,** 1–12 (2005).
23. Mo, Y. Y., Yu, Y., Ee, P. L. & Beck, W. T. Overexpression of a dominant-negative mutant Ubc9 is associated with increased sensitivity to anticancer drugs. *Cancer Res.* **64,** 2793–2798 (2004).
24. Zhao, X. & Blobel, G. A. SUMO ligase is part of a nuclear multiprotein complex that affects DNA repair and chromosomal organization. *Proc. Natl Acad. Sci. USA* **102,** 4777–4782 (2005).
25. Potts, P. R. & Yu, H. Human MMS21/NSE2 is a SUMO ligase required for DNA repair. *Mol. Cell. Biol.* **25,** 7021–7032 (2005).
26. Mabb, A. M., Wuerzberger-Davis, S. M. & Miyamoto, S. PIASy mediates NEMO sumoylation and NF-κB activation in response to genotoxic stress. *Nature Cell Biol.* **8,** 986–993 (2006).
27. Ishiai, M. *et al.* DNA cross-link repair protein SNM1A interacts with PIAS1 in nuclear focus formation. *Mol. Cell. Biol.* **24,** 10733–10741 (2004).
28. Boulton, S. J. *et al.* BRCA1/BARD1 orthologs required for DNA repair in *Caenorhabditis elegans. Curr. Biol.* **14,** 33–39 (2004).
29. Park, M. A., Seok, Y. J., Jeong, G. & Lee, J. S. SUMO1 negatively regulates BRCA1-mediated transcription, via modulation of promoter occupancy. *Nucleic Acids Res.* **36,** 263–283 (2008).
30. Peter, M. & Ameer-Beg, S. M. Imaging molecular interactions by multiphoton FLIM. *Biol. Cell* **96,** 231–236 (2004).
31. Peter, M. *et al.* Multiphoton-FLIM quantification of the EGFP-mRFP1 FRET pair for localization of membrane receptor-kinase interactions. *Biophys. J.* **88,** 1224–1237 (2005).
32. Ng, T. *et al.* Imaging protein kinase Cα activation in cells. *Science* **283,** 2085–2089 (1999).
33. Ganesan, S., Ameer-Beg, S. M., Ng, T. T., Vojnovic, B. & Wouters, F. S. A dark yellow fluorescent protein (YFP)-based resonance energy-accepting chromoprotein (REACh) for Forster resonance energy transfer with GFP. *Proc. Natl Acad. Sci. USA* **103,** 4089–4094 (2006).
34. Dadke, S. *et al.* Regulation of protein tyrosine phosphatase 1B by sumoylation. *Nature Cell Biol.* **9,** 80–85 (2007).
35. Bossis, G. & Melchior, F. Regulation of SUMOylation by reversible oxidation of SUMO conjugating enzymes. *Mol. Cell* **21,** 349–357 (2006).
36. Schmidt, D. & Muller, S. PIAS/SUMO: new partners in transcriptional regulation. *Cell. Mol. Life Sci.* **60,** 2561–2574 (2003).
37. Munarriz, E. *et al.* PIAS-1 is a checkpoint regulator which affects exit from G1 and G2 by sumoylation of p73. *Mol. Cell. Biol.* **24,** 10593–10610 (2004).
38. Kim, H. & Chen, J. New players in the BRCA1-mediated DNA damage responsive pathway. *Mol. Cells* **25,** 457–461 (2008).
39. Nishikawa, H., Ooka, S., Sato, K., Arima, K., Okamoto, J., Klevit, R. E., Fukuda, M. & Ohta, T. Mass spectrometric and mutational analyses reveal Lys-6-linked polyubiquitin chains catalyzed by BRCA1-BARD1 ubiquitin ligase. *J Biol Chem,* (2003).
40. Wu-Baer, F., Lagrazon, K., Yuan, W. & Baer, R. The BRCA1/BARD1 heterodimer assembles polyubiquitin chains through an unconventional linkage involving lysine residue K6 of ubiquitin. *J. Biol. Chem.* **278,** 34743–34746 (2003).
41. Mallery, D. L., Vandenberg, C. J. & Hiom, K. Activation of the E3 ligase function of the BRCA1/BARD1 complex by polyubiquitin chains. *EMBO J.* **21,** 6755–6762 (2002).
42. Christensen, D. E., Brzovic, P. S. & Klevit, R. E. E2-BRCA1 RING interactions dictate synthesis of mono- or specific polyubiquitin chain linkages. *Nature Struct. Mol. Biol.* **14,** 941–948 (2007).
43. Manke, I. A., Lowery, D. M., Nguyen, A. & Yaffe, M. B. BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science* **302,** 636–639 (2003).
44. Jaffray, E. G. & Hay, R. T. Detection of modification by ubiquitin-like proteins. *Methods* **38,** 35–38 (2006).
45. Prag, S. *et al.* Activated ezrin promotes cell migration through recruitment of the GEF Dbl to lipid rafts and preferential downstream activation of Cdc42. *Mol. Biol. Cell* **18,** 2935–2948 (2007).
46. Barber, P. R., Ameer-Beg, S. M., Gilbey, J. D., Edens, R. J., Ezike, I. & Vojnovic, B. Global and pixel kinetic data analysis for FRET detection by multi-photon time-domain FLIM. *Proc. SPIE* **5700,** 171–181 (2005).
47. Bennardo, N., Cheng, A., Huang, N. & Stark, J. M. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. *PLoS Genet.* **4,** e1000110 (2008).
48. Boutell, C., Orr, A. & Everett, R. D. PML residue lysine 160 is required for the degradation of PML induced by herpes simplex virus type 1 regulatory protein ICP0. *J. Virol.* **77,** 8686–8694 (2003).
49. Boutell, C., Sadis, S. & Everett, R. D. Herpes simplex virus type 1 immediate-early protein ICP0 and its isolated RING finger domain act as ubiquitin E3 ligases *in vitro. J. Virol.* **76,** 841–850 (2002).
50. Bliss, C. I. The toxicity of poisons applied jointly. *Ann. Appl. Biol.* **26,** 585–615 (1939).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We are grateful to V. De Laurenzi for the PIAS1 expression constructs, to R. Hay for His-SUMO1 and His-SUMO2 cells, to D. Durocher for anti-RNF168 antibody, to G. Stewart for discussions and S. Jackson for sharing results before publication. The work was supported by grants from Breast Cancer Campaign (to J.R.M., A.A., #SF06, and L.B., #06NovPHD13Morris), Cancer Research UK (to R.D., #C8820/A9494), the Medical Research Council (to E.S. and D.W., #6900577, and C.B.), the Richard Dimbleby Cancer Fund to King's College London (to M.K. and T.N.) and Breakthrough Breast Cancer (to T.K). Multiphoton FLIM systems and acquisition/analysis software were built by S. Ameer-Beg, P. Barber and B. Vojnovic, with support from MRC Co-operative Group Grant G0100152 #56891 and UK Research Councils Basic Technology Research Programme Grant GR/R87901/01.

**Author Contributions** J.R.M. conceived and designed the study, generated reagents, performed experiments and wrote the paper. C.B. performed *in vitro* assays, confirmed and developed the initial concept, and generated reagents. M.K. optimised and performed FLIM measurements and analysis. R.D. and D.W. performed experiments and generated reagents. L.B. performed co-localisation observations, A.A. and L.P. generated reagents and Y.G. generated reagents and participated in discussions. T.K. undertook FLIM measurements. T.N. provided expertise and input into the design of the FLIM experiments, and E.S. provided advice and mentoring to J.R.M.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.R.M. (jo.morris@genetics.kcl.ac.uk).

## METHODS

**Plasmid constructs.** Full-length BRCA1, SUMO1 and SUMO1-GA were cloned 3′ of the modified RFP of pcDNA3.1 (ref. 31) or in myc-pcDNA3.1, and full-length BRCA1 and UBC9 were cloned 3′ of the GFP in p-EGFP (Clontech). Flag–HIS-MMS21 complementary DNA was cloned into pCL-NCX previously modified to contain 3×Flag–HIS tag, and PIAS4 in 3×Flag–Stag in pCDNA3.1(−) modified to contain 3Flag–Stag. PIAS1 expression constructs were gifts (see Acknowledgements). Point mutations were generated using site-directed mutagensis and confirmed by sequencing. Ubiquitin and BARD1 clones in pcDNA3.1 have been previously described[5].

**Cell treatments.** Cells were treated with 3 mM hydroxyurea for 8 h, 0.9% NaCl carrier, 10 μM cisplatin for 3 h followed by 16 h recovery, or 16 h in 20 nM epirubicin, 3 mM hydroxyurea for 8 h followed by 1 mM $H_2O_2$ for 15 min or 15 min heat shock at 43 °C or cells were exposed to 10 Gy irradiation using a Gammacell 1000 Elite irradiator (caesium-137 source).

**siRNA.** On-target Plus SMART Pool siRNAs (Dharmacon) used were: BRCA1 (L-003461-00), PIAS1 (L-008167-00), PIAS2 (L-009428-00), PIAS3 (L-004164-00) PIAS4 (L-006445-00) and MMS21/NSE2 (L-018070-00), RNF168 (L007152-00-0005), RNF8 (L-009600-00-0005). Non-targeting control siRNA has been confirmed to have minimal targeting of known genes (D-001810-10-05). siRNAs to untranslated regions used were to target sequence: GGCGAAG UUCACUGCGC (PIAS1), CAGAGGGAGGAGUGACC (PIAS4). Knockdowns were confirmed by RT–PCR of extracted cell-line RNA and were designed over more than one exon boundary to encompass the active site (RING).

**Purification of 6×His tagged SUMO conjugates.** His-SUMO stable Hela cells[44] were maintained in 0.5 μg ml$^{-1}$ puromycin and were a gift from R. Hay. Cells were lysed directly in 8 M urea, 0.1 M $Na_2HPO_4/NaH_2PO_4$, 0.01 M Tris–HCl, pH 6.3, 10 mM β-mercaptoethanol, 5 mM imidazole plus 0.2% Triton-X-100, harvested and sonicated. They were then mixed with 50 μl of $Ni^{2+}$ Talon agarose beads (BD-Bioscience) and incubated overnight at 4 °C, washed and eluted in SDS–polyacrylamide gel electrophoresis buffer according to ref. 44.

**Immunofluorescence microscopy.** All cells were grown and prepared as previously described[5].

**Antibodies.** The antibodies used in the study were MS110 ascites (Ab1, Calbiochem), FK-2 (Biomol), anti-BARD1 antibody (Serotec), anti-Flag (M2) (Sigma), control rabbit IgG (Sigma), β-actin (Abcam), anti-c-myc (9E10), anti-SUMO1, anti-SUMO2/3, (Santa Cruz), γ-H2AX clone JBW301 (Millipore) and polyclonal anti-γ-H2AX (Abcam), anti-RNF8 (Abnova), anti-RNF168 (ref. 10), anti-RAP80 (Bethyl) and anti-K63-ubiquitin (Millipore).

**Time-resolved multiphoton microscopy.** FLIM measurements were undertaken with a modified multiphoton microscopy system similar to that described previously[45]. Fluorescence lifetime imaging capability was provided by time-correlated single-photon counting electronics (Becker & Hickl, SPC 830). Data were collected through a bandpass filter centred at $\lambda = 510 \pm 10$nm (Chroma). Excitation power was adjusted using a neutral-density filter to give average photon counting rates of the order $10^4$–$10^5$ photons s$^{-1}$ to avoid pulse pile up. Acquisition times of the order of 300 s at low excitation power were used to achieve sufficient photon statistics for fitting, while avoiding either pulse pile up or observable photo-bleaching. The imaging system was controlled, and the data later analysed, with custom software written in CVI LabWindows[46]. FRET efficiency $= 1 - \tau da/\tau di$, where d*a* is the pixel-by-pixel fluorescence lifetime of the donor in the presence of acceptor and d*i* is the average lifetime of the donor in the absence of acceptor (in all experiments unlabelled BARD1 was co-expressed). The Förster radius (distance at which the efficiency of energy transfer is 50%) of the GFP and RFP pair has been calculated as 4.7 nm (ref. 31). Note that in analysis of FRET data there are two elements that must be considered: interacting fluorophore population and FRET efficiency. Bulk measurements of FRET efficiency cannot distinguish between an increase in FRET efficiency (that is, proximity) and an increase in FRET population (concentration of interacting species) because the two parameters are not resolved.

**Repair assays.** HeLa clones with the homologous recombination (DR–GFP) and total non-homologous end-joining (EJ5–GFP) reporters stably integrated were generated as previously described[47]. For repair assays, HeLa-DR–GFP or Hela-EJ5–GFP were either mock transfected or transfected with non-targeting or targeting siRNA. Cells were left for 24 h before transfection with the I-sceI expression vector pCBA-I-sceI. Three days after pCBA-I-sceI transfection, cells were fixed and the proportion of GFP-positive cells counted. Counts were performed in triplicate.

**Clonogenic cell survival assays.** 293T cells were plated onto 24-well tissue culture dishes at ($10^5$ cells per well) and transfected with siRNA and plasmid with Dharmafect according to the manufacturer's instructions. After 3 days, they were exposed to cisplatin (Sigma Chemicals) for 3 h and replated into 10-cm dishes at various concentrations. After 11 days, cells were fixed with methanol for 10 min and stained with 0.5% crystal violet (BDH Chemicals). Washed dishes were dried, and colonies >1 mm were scored. Half-maximum inhibitory concentration ($IC_{50}$) values were calculated for each siRNA from the respective sigmoidal dose–response curves using Prism software. The Bliss independence model is defined by the equation $Exy = Ex + Ey - (ExEy)$, where *Exy* is the additive effect of siRNA 1 and 2 as predicted by their individual effects (*Ex* and *Ey*)[50].

**Protein production, SUMO conjugation and ubiquitin ligase assays.** Bacterial expression of human BRCA1/BARD1 heterodimer was from bi-cistronic expression vector purified using nickel resin as described previously[4].

***In vitro* SUMO conjugation assays.** These were performed as described by Boutell *et al.*[48,49]. Purification of SUMOylated complexes is described in Supplementary Fig. 2.

**Immunoprecipitation.** 293T cells were lysed in 20 mM Tris-HCL, pH 8, 137 mM NaCl, 1 mM EGTA, 1% Triton-X-100, 10% glycerol, 1.5 $MgCl_2$, containing 10 mM iodoactetamide and protease inhibitors. After clearance, lysate was incubated with 20 μl MS110 (Ab1) overnight. Beads were washed in lysis buffer before elution in SDS–polyacrylamide gel electrophoresis buffer.

LETTERS

# A super-Earth transiting a nearby low-mass star

David Charbonneau[1], Zachory K. Berta[1], Jonathan Irwin[1], Christopher J. Burke[1], Philip Nutzman[1], Lars A. Buchhave[1,2], Christophe Lovis[3], Xavier Bonfils[3,4], David W. Latham[1], Stéphane Udry[3], Ruth A. Murray-Clay[1], Matthew J. Holman[1], Emilio E. Falco[1], Joshua N. Winn[5], Didier Queloz[3], Francesco Pepe[3], Michel Mayor[3], Xavier Delfosse[4] & Thierry Forveille[4]

**A decade ago, the detection of the first[1,2] transiting extrasolar planet provided a direct constraint on its composition and opened the door to spectroscopic investigations of extrasolar planetary atmospheres[3]. Because such characterization studies are feasible only for transiting systems that are both nearby and for which the planet-to-star radius ratio is relatively large, nearby small stars have been surveyed intensively. Doppler studies[4–6] and microlensing[7] have uncovered a population of planets with minimum masses of 1.9–10 times the Earth's mass ($M_\oplus$), called super-Earths. The first constraint on the bulk composition of this novel class of planets was afforded by CoRoT-7b (refs 8, 9), but the distance and size of its star preclude atmospheric studies in the foreseeable future. Here we report observations of the transiting planet GJ 1214b, which has a mass of 6.55$M_\oplus$ and a radius 2.68 times Earth's radius ($R_\oplus$), indicating that it is intermediate in stature between Earth and the ice giants of the Solar System. We find that the planetary mass and radius are consistent with a composition of primarily water enshrouded by a hydrogen–helium envelope that is only 0.05% of the mass of the planet. The atmosphere is probably escaping hydrodynamically, indicating that it has undergone significant evolution during its history. The star is small and only 13 parsecs away, so the planetary atmosphere is amenable to study with current observatories.**

The recently commissioned MEarth Project[10,11] uses an array of eight identical 40-cm automated telescopes to photometrically monitor 2,000 nearby M dwarfs with masses between 0.10 and 0.35 solar masses ($M_\odot$) drawn from a sample[12] of nearby stars with a large proper motion. After applying a trend-filtering algorithm[13] and a three-day running median filter to remove long-term stellar variability, we searched[14] the light curves for evidence of periodic eclipsing signals. The light curve of the star GJ 1214 contained 225 data points, of which six values were consistent with having been gathered during a time of eclipse and indicating a signal with a period of 1.58 days. On the basis of this prediction, we gathered additional photometric observations at high cadence using the eight telescopes of the MEarth array as well as the adjacent 1.2-m telescope. These light curves (shown in Fig. 1) confirm that the star is undergoing flat-bottomed eclipses with a depth of 1.3%, indicative of a planetary transit. Astrophysical false positives that result from blends of eclipsing binary stars and hinder field transit surveys are not[10,11] a concern under the strategy of the MEarth survey. GJ 1214 has a large proper motion, and by examining archival images we established that no second star lies at the current sky position of GJ 1214, ruling out a blend resulting from an eclipsing binary that is not physically associated with the target. The measured parallax and photometry of GJ 1214 (Table 1) place stringent constraints on the presence of an unresolved physically associated binary companion: we find no physically plausible coeval model that matches both the observed transit depth and the short duration of ingress and egress. We subsequently used the HARPS[5,6] instrument to gather radial velocity observations (Fig. 2 and Supplementary Information), which confirmed the planetary nature of the companion and permitted us to estimate its mass.

Table 1 presents our estimates of the physical quantities for planet and star. We estimate the planetary equilibrium temperature to be as great as 555 K (the case for a Bond albedo of 0) and as low as 393 K (assuming a Bond albedo of 0.75, the same as that for Venus). This latter value is significantly cooler than all known transiting planets, and exceeds the condensation point of water by only 20 K. This consideration is significant, because it demonstrates that for M dwarfs the discovery of super-Earths within the stellar habitable zones is within reach of ground-based observatories such as MEarth,

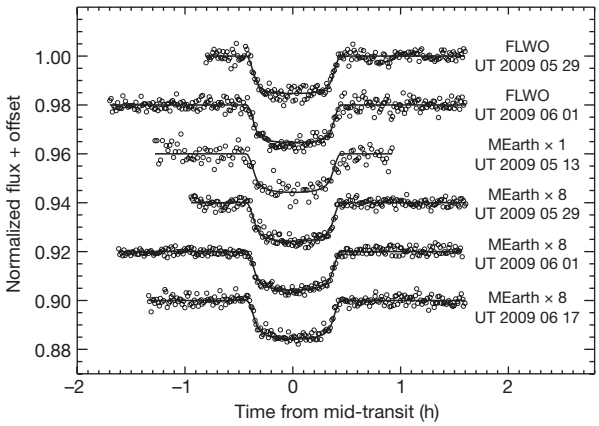



**Figure 1 | Photometric data for GJ 1214.** Light curves of GJ 1214 spanning times of transit for four separate transit events, gathered with the MEarth Observatory (either a single telescope or eight telescopes, denoted respectively by MEarth × 1 and MEarth × 8) and the F. L. Whipple Observatory (FLWO) 1.2-m telescope. All light curves have been binned to a uniform cadence of 45 s to facilitate a visual comparison. We fitted the unbinned light curves to a model[29] corresponding to a planet in a circular orbit transiting a limb-darkened star, setting the limb-darkening coefficients to match the inferred stellar properties as described in the text. This model has five parameters: the orbital period $P$, the time of transit centre $T_c$, the ratio of the radius of the planet to that of the star $R_p/R_s$, the ratio of the semimajor axis to the stellar radius $a/R_s$, and the orbital inclination $i$. We use a Markov chain Monte Carlo method to estimate the uncertainties, and our results are stated in Table 1. The solid lines show the best-fit model fitted simultaneously to all the data.

[1]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138 USA. [2]Niels Bohr Institute, Copenhagen University, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark. [3]Observatoire de Genève, Université de Genève, 51 chemin des Maillettes, 1290 Sauverny, Switzerland. [4]Université Joseph Fourier – Grenoble 1, Centre national de la recherche scientifique, Laboratoire d'Astrophysique de Grenoble (LAOG), UMR 5571, 38041 Grenoble Cedex 09, France. [5]Department of Physics, Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

**Table 1 | System parameters for GJ 1214**

| Parameter | Value |
|---|---|
| Orbital period, $P$ (days) | $1.5803925 \pm 0.0000117$ |
| Times of centre of transit, $T_c$ (HJD) | $2454964.944208 \pm 0.000403$<br>$2454980.7479702 \pm 0.0000903$<br>$2454983.9087558 \pm 0.0000901$<br>$2454999.712703 \pm 0.000126$ |
| Planet/star radius ratio, $R_p/R_s$ | $0.1162 \pm 0.00067$ |
| Scaled semimajor axis, $a/R_s$ | $14.66 \pm 0.41$ |
| Impact parameter, $b$ | $0.354^{+0.061}_{-0.082}$ |
| Orbital inclination, $i$ (deg) | $88.62^{+0.35}_{-0.28}$ |
| Radial velocity semi-amplitude, $K$ (m s$^{-1}$) | $12.2 \pm 1.6$ |
| Systemic velocity, $\gamma$ (m s$^{-1}$) | $-21{,}100 \pm 1{,}000$ |
| Orbital eccentricity, $e$ | $<0.27$ (95% confidence) |
| Stellar mass, $M_s$ | $0.157 \pm 0.019 M_\odot$ |
| Stellar radius, $R_s$ | $0.2110 \pm 0.0097 R_\odot$ |
| Stellar density, $\rho_s$ (kg m$^{-3}$) | $23{,}900 \pm 2{,}100$ |
| Log of stellar surface gravity (CGS units), $\log g_s$ | $4.991 \pm 0.029$ |
| Stellar projected rotational velocity, $v \sin i$ (km s$^{-1}$) | $<2.0$ |
| Stellar parallax (mas) | $77.2 \pm 5.4$ |
| Stellar photometry | |
| $V$ | $15.1 \pm 0.6$ |
| $I$ | $11.52 \pm 0.1$ |
| $J$ | $9.750 \pm 0.024$ |
| $H$ | $9.094 \pm 0.024$ |
| $K$ | $8.782 \pm 0.020$ |
| Stellar luminosity, $L_s$ | $0.00328 \pm 0.00045 L_\odot$ |
| Stellar effective temperature, $T_{eff}$ (K) | $3{,}026 \pm 130$ |
| Planetary radius, $R_p$ | $2.678 \pm 0.13 R_\oplus$ |
| Planetary mass, $M_p$ | $6.55 \pm 0.98 M_\oplus$ |
| Planetary density, $\rho_p$ (kg m$^{-3}$) | $1870 \pm 400$ |
| Planetary surface acceleration under gravity, $g_p$ (m s$^{-2}$) | $8.93 \pm 1.3$ |
| Planetary equilibrium temperature, $T_{eq}$ (K) | |
| Assuming a Bond albedo of 0 | 555 |
| Assuming a Bond albedo of 0.75 | 393 |

To convert the photometric and radial velocity parameters into physical parameters for the system, we require a constraint on the stellar mass. Using the observed parallax distance[26] of $12.95 \pm 0.9$ pc and apparent $K$-band brightness, we employ an empirical relation[27] between stellar mass and absolute $K$-band magnitude to estimate the stellar mass. With this value we find the planetary radius and mass. The uncertainty on the planet mass is the quadrature sum of the propagated uncertainties on the radial-velocity amplitude and those from the uncertainty in the stellar mass, which contribute $0.85 M_\oplus$ and $0.50 M_\oplus$ to the error budget, respectively. We use the observed $I-K$ colour and an empirical relation[28] to estimate the bolometric correction and subsequently the stellar luminosity and stellar effective temperature (assuming the stellar radius quoted in the table). Using the luminosity, we estimate a planetary equilibrium temperature, assuming a value for the Bond albedo. HJD, heliocentric Julian date.

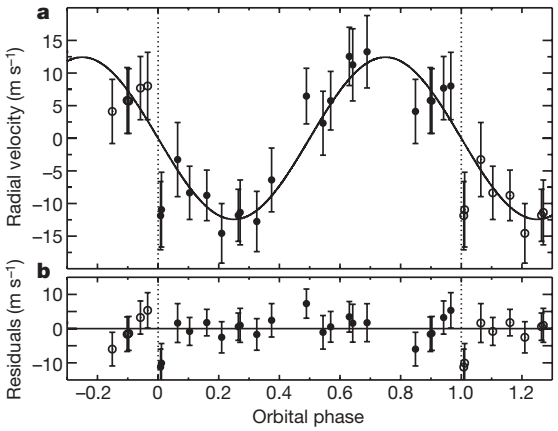



**Figure 2 | Change in radial velocity of GJ 1214. a**, We gathered 21 observations during 2009 July 24 to 2009 August 6, and six observations during 2009 June 11–19. We estimate[30] the change in the radial velocity by first constructing a stellar template by summing the observations (corrected to the barycentre), and then minimizing the $\chi^2$ difference between this template and each spectrum. We initially restricted our analysis to the July–August data (shown as filled points, with repetitions shown as open symbols), out of concern that long-term stellar variability or a second planet could lead to an offset between these data and those gathered in June (not shown). We fitted a sinusoidal model (solid curve) constrained by the photometric period and time of transit (dotted lines) and found a good fit ($\chi^2 = 15.98$ for 19 degrees of freedom) with a semi-amplitude of $K = 12.2 \pm 1.6$ m s$^{-1}$. We considered an eccentric orbit, and found that the best-fit model ($\chi^2 = 13.02$ for 17 degrees of freedom) was not significantly better and yielded an indistinguishable $K$. We conclude that there is no evidence that the orbit is non-circular, and we state the upper limit in Table 1. We then included the June observations and found $K = 12.4 \pm 1.8$ m s$^{-1}$, which is consistent with but noisier than the previous estimate. However, to obtain a $\chi^2$ consistent with an acceptable fit, we need to introduce an additional noise term of 2.7 m s$^{-1}$, or an offset of $-8$ m s$^{-1}$ from the June data to the July–August data. Our photometry indicates that the stellar brightness varies by 2% on timescales of several weeks. We conclude that spot-induced stellar jitter is the most likely explanation. **b**, Residuals of the July–August data to the sinusoidal model. The residuals are consistent with the internal estimates of the uncertainties, shown here as $1\sigma$ error bars.

whereas the discovery of such objects orbiting solar analogues is thought to require space-based platforms such as the Kepler Mission[15].

We compare in Fig. 3 the measured mass and radius of GJ 1214b with that of models[16] that predict planetary radii as a function of mass and assumed composition. We consider a hypothetical[16] water-dominated composition (75% $H_2O$, 22% Si and 3% Fe) and take this prediction to be an upper bound on the planet radius, assuming a solid composition. This model provides a minimum mass for the gaseous envelope: assuming that the envelope is isothermal (with a temperature corresponding to a Bond albedo of 0, above) and composed of hydrogen and helium, and that the observed planetary transit radius corresponds[17] to an atmospheric pressure of 1 mbar, we estimate a scale height of 233 km and a total envelope mass of $0.0032 M_\oplus$ (0.05% of the planetary mass). In this model, the relative mass of the envelope to the core is much smaller than that for the ice giants of the Solar System. If we continue under this assumed composition and consider both the Solar System planets and the extrasolar worlds together in Fig. 3, the sequence decreasing in mass from HD 149026b and Saturn to HAT-P-11b, GJ 436b, Neptune and Uranus, and finally GJ 1214b would then trace an atmospheric depletion curve: the mass of the gaseous envelope relative to that of the core would decrease with mass, which is consistent with the fact that the atmospheres of Earth and Venus are each only a trace component by mass. We note, however, that with only an estimate of the average density, we cannot be certain that GJ 1214b, GJ 436b and HAT-P-11b do not have compositions significantly different from that assumed above. For example, these planets could contain cores of iron or silicates enshrouded by much more massive envelopes of hydrogen and helium, a situation that would challenge models of formation but is not excluded by the current observations.

Our estimate of the stellar radius is 15% larger than that predicted by theoretical models[18] for the stellar mass we derived. Such discrepancies are well established from observations of M-dwarf eclipsing binaries, and indeed a similar stellar radius enhancement was determined[19] for the only other M-dwarf with a known transiting planet, GJ 436. If the true value of the stellar radius is $0.18 R_\odot$ (as predicted by both the theoretical models[18] and an empirical radius relation[20] for low-mass stars), then the planet radius would be revised downwards to $2.27 R_\oplus$, which is consistent with a water-dominated composition without the need for a gaseous envelope. If the empirical relation[21] for angular diameter can be extended to this spectral type, this would provide an alternative estimate of the stellar radius, given a refined estimate of the parallax.

We considered the timescale for hydrodynamic escape of a hydrogen-dominated envelope. Assuming that the ultraviolet luminosity of the star is $10^{-5}$ of its bolometric luminosity (typical[22] for inactive field M dwarfs), we calculate[23] a hydrodynamical escape rate of $9 \times 10^5$ kg s$^{-1}$; we further verified that at the sonic point the mean free path is only 4% of the scale height. At this rate, the minimum-mass envelope described above would be removed in about 700 Myr. The stellar ultraviolet radiation was probably much larger when the star was young, which would result in an even shorter timescale for

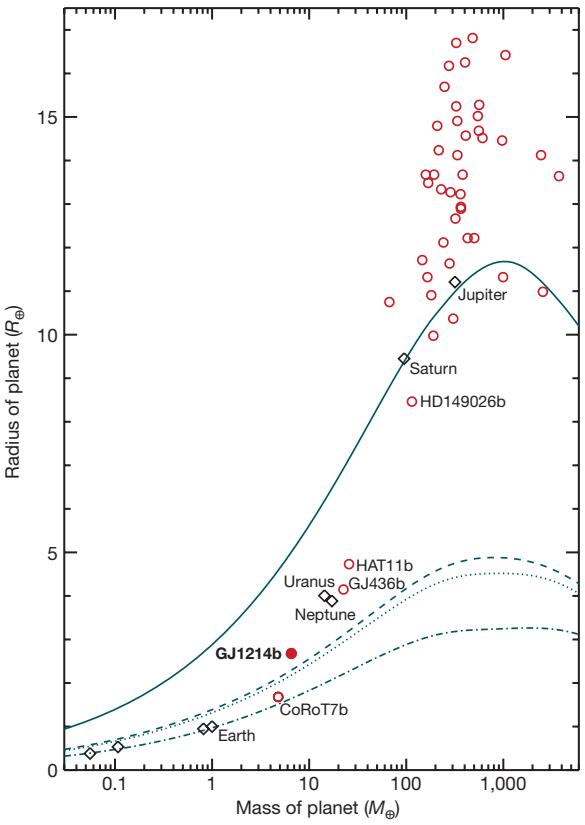



**Figure 3 | Masses and radii of transiting planets.** GJ 1214b is shown as a red filled circle (the 1$\sigma$ uncertainties correspond to the size of the symbol), and the other known transiting planets are shown as open red circles. The eight planets of the Solar System are shown as black diamonds. GJ 1214b and CoRoT-7b are the only extrasolar planets with both well-determined masses and radii for which the values are less than those for the ice giants of the Solar System. Despite their indistinguishable masses, these two planets probably have very different compositions. Predicted[16] radii as a function of mass are shown for assumed compositions of H/He (solid line), pure $H_2O$ (dashed line), a hypothetical[16] water-dominated world (75% $H_2O$, 22% Si and 3% Fe core; dotted line) and Earth-like (67.5% Si mantle and a 32.5% Fe core; dot-dashed line). The radius of GJ 1214b lies $0.49 \pm 0.13$ $R_\oplus$ above the water-world curve, indicating that even if the planet is predominantly water in composition, it probably has a substantial gaseous envelope.

removal of the envelope. An age of 3–10 Gyr for the star is supported[24] both by its kinematics (which indicate that it is a member of the old disk) and the lack of chromospheric activity from the absence of H$\alpha$ line emission. Moreover, the dominant periodicity in the MEarth photometry is 83 days. Stars spin down as they age, and a very long rotation would also indicate an old star. Thus we conclude that significant loss of atmospheric mass has occurred over the lifetime of the planet; the current envelope is therefore probably not primordial. Moreover, some (or all) of the present envelope may have resulted from outgassing and further photodissociation of material from the core. If the composition of the gaseous envelope is indeed dominated by hydrogen (whether primordial or not), the annulus of the transmissive portion of planetary atmosphere would occult roughly 0.16% of the stellar disk during transit and thus present a signal larger than that already studied for other exoplanets[3]. Thus GJ 1214b presents an opportunity to study a non-primordial atmosphere enshrouding a world orbiting another star. Such studies have been awaited[25] and would serve to confirm directly that the atmosphere was predominantly hydrogen, because only then would the scale height be large enough to present a measurable wavelength-dependent signal in transit.

The discussion above assumes that the solid core of GJ 1214b is predominantly water. This is at odds with the recently discovered[8,9] CoRoT-7b, the only other known transiting super-Earth. CoRot-7b has mass of $4.8M_\oplus$, a radius of $1.7R_\oplus$ and a density of 5,600 kg m$^{-3}$, indicating a composition that is predominantly rock. The very different radii of GJ 1214b and CoRoT-7b despite their indistinguishable masses may be related to the differing degrees to which the two planets are irradiated by their parent stars: owing to the much greater luminosity of its central star, CoRoT-7b has an equilibrium temperature of about 2,000 K, roughly fourfold that of GJ 1214b. It may be that both planets have rocky cores of similar mass and that it is only for CoRoT-7b that the gaseous envelope has been removed, yielding the smaller observed radius. Alternatively, GJ 1214b may have a water-dominated core, indicating a very different formation history from that of CoRoT-7b. Such degeneracies in the models[16] of the physical structures of super-Earths will be commonplace when only a radius and mass are available, but at least one method[25] has been proposed to mitigate this problem in part. The differences in composition between GJ 1214b and CoRoT-7b bear on the quest for habitable worlds: numerous planets with masses indistinguishable from those of GJ 1214b and CoRoT-7b have been uncovered indirectly by radial velocity studies, and some of these lie in or near their stellar habitable zones. If such cooler super-Earth planets do indeed have gaseous envelopes similar to that of GJ 1214b, the extreme atmospheric pressure and absence of stellar radiation at the surface might render them inhospitable to life as we know it on Earth. This would motivate the push to even more sensitive ground-based techniques capable of detecting planets with sizes and masses equal to that of the Earth orbiting within the habitable zones of low-mass stars.

**Received 20 October; accepted 17 November 2009.**


1. Charbonneau, D., Brown, T. M., Latham, D. W. & Mayor, M. Detection of planetary transits across a sun-like star. *Astrophys. J.* **529,** L45–L48 (2000).
2. Henry, G. W., Marcy, G. W., Butler, R. P. & Vogt, S. S. A transiting '51-Peg-like' planet. *Astrophys. J.* **529,** L41–L44 (2000).
3. Charbonneau, D., Brown, T. M., Burrows, A. & Laughlin, G. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) 701–716 (Univ. Arizona Press, 2007).
4. Rivera, E. J. *et al.* A $\sim$7.5 $M_\oplus$ planet orbiting the nearby star, GJ 876. *Astrophys. J.* **634,** 625–640 (2005).
5. Udry, S. *et al.* The HARPS search for southern extra-solar planets. XI. Super-Earths (5 and 8 $M_\oplus$) in a 3-planet system. *Astron. Astrophys.* **469,** L43–L47 (2007).
6. Mayor, M. *et al.* The HARPS search for southern extra-solar planets. XVIII. An Earth-mass planet in the GJ 581 planetary system. *Astron. Astrophys.* **507,** 487–494 (2009).
7. Beaulieu, J.-P. *et al.* Discovery of a cool planet of 5.5 Earth masses through gravitational microlensing. *Nature* **439,** 437–440 (2006).
8. Léger, A. *et al.* Transiting exoplanets from the CoRoT space mission. VIII. CoRoT-7b: the first super-Earth with a measured radius. *Astron. Astrophys.* **506,** 287–302 (2009).
9. Queloz, D. *et al.* The CoRoT-7 planetary system: two orbiting super-Earths. *Astron. Astrophys.* **506,** 303–319 (2009).
10. Nutzman, P. & Charbonneau, D. Design considerations for a ground-based transit search for habitable planets orbiting M dwarfs. *Publ. Astron. Soc. Pacif.* **120,** 317–327 (2008).
11. Irwin, J. *et al.* GJ 3236: A new bright, very low mass eclipsing binary system discovered by the MEarth Observatory. *Astrophys. J.* **701,** 1436–1449 (2009).
12. Lépine, S. Nearby stars from the LSPM-North Proper-Motion Catalog. I. main-sequence dwarfs and giants within 33 parsecs of the sun. *Astron. J.* **130,** 1680–1692 (2005).
13. Kovács, G., Bakos, G. & Noyes, R. W. A trend filtering algorithm for wide-field variability surveys. *Mon. Not. R. Astron. Soc.* **356,** 557–567 (2005).
14. Burke, C. J., Gaudi, B. S., DePoy, D. L. & Pogge, R. W. Survey for transiting extrasolar planets in stellar systems. III. A limit on the fraction of stars with planets in the open cluster NGC 1245. *Astron. J.* **132,** 210–230 (2006).
15. Borucki, W. J. *et al.* Kepler's optical phase curve of the exoplanet HAT-P-7b. *Science* **325,** 709 (2009).
16. Seager, S., Kuchner, M., Hier-Majumder, C. A. & Militzer, B. Mass–radius relationships for solid exoplanets. *Astrophys. J.* **669,** 1279–1297 (2007).
17. Burrows, A., Sudarsky, D. & Hubbard, W. B. A theory for the radius of the transiting giant planet HD 209458b. *Astrophys. J.* **594,** 545–551 (2003).
18. Baraffe, I., Chabrier, G., Allard, F. & Hauschildt, P. H. Evolutionary models for solar metallicity low-mass stars: mass–magnitude relationships and color–magnitude diagrams. *Astron. Astrophys.* **337,** 403–412 (1998).
19. Torres, G. The transiting exoplanet host star GJ 436: a test of stellar evolution models in the lower main sequence, and revised planetary parameters. *Astrophys. J.* **671,** L65–L68 (2007).
20. Demory, B.-O. *et al.* Mass–radius relation of low and very low-mass stars revisited with the VLTI. *Astron. Astrophys.* **505,** 205–215 (2009).

21. Kervella, P., Thévenin, F., Di Folco, E. & Ségransan, D. The angular sizes of dwarf stars and subgiants. Surface brightness relations calibrated by interferometry. *Astron. Astrophys.* **426,** 297–307 (2004).
22. Walkowicz, L. M., Johns-Krull, C. M. & Hawley, S. L. Characterizing the near-UV environment of M dwarfs. *Astrophys. J.* **677,** 593–606 (2008).
23. Murray-Clay, R. A., Chiang, E. I. & Murray, N. Atmospheric escape from hot Jupiters. *Astrophys. J.* **693,** 23–42 (2009).
24. Reid, I. N., Hawley, S. L. & Gizis, J. E. The Palomar/MSU nearby-star spectroscopic survey. I. The northern M dwarfs—bandstrengths and kinematics. *Astron. J.* **110,** 1838–1859 (1995).
25. Miller-Ricci, E., Seager, S. & Sasselov, D. The atmospheric signatures of super-Earths: how to distinguish between hydrogen-rich and hydrogen-poor atmospheres. *Astrophys. J.* **690,** 1056–1067 (2009).
26. van Altena, W. F., Lee, J. T. & Hoffleit, E. D. *The General Catalogue of Trigonometric Stellar Parallaxes* 4th edn (Yale Univ. Observatory, 1995).
27. Delfosse, X. *et al.* Accurate masses of very low mass stars. IV. Improved mass–luminosity relations. *Astron. Astrophys.* **364,** 217–224 (2000).
28. Leggett, S. K. *et al.* Spectral energy distributions for disk and halo M dwarfs. *Astrophys. J.* **535,** 965–974 (2000).
29. Mandel, K. & Agol, E. Analytic light curves for planetary transit searches. *Astrophys. J.* **580,** L171–L175 (2002).
30. Bouchy, F., Pepe, F. & Queloz, D. Fundamental photon noise limit to radial velocity measurements. *Astron. Astrophys.* **374,** 733–739 (2001).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank M. Everett for gathering the FLWO 1.2-m observations, S. Seager for providing a digital version of the structural models, and D. Sasselov and S. Seager for comments on the manuscript. Support for this work was provided by the David and Lucile Packard Foundation Fellowship for Science and Engineering awarded to D.C., and by the US National Science Foundation under grant number AST-0807690. L.A.B. and D.W.L. acknowledge support from the NASA Kepler mission under cooperative agreement NCC2-1390. M.J.H. acknowledges support by NASA Origins Grant NNX09AB33G. The HARPS observations were gathered under the European Southern Observatory Director's Discretionary Program 283.C-5022 (A). We thank the Smithsonian Astrophysical Observatory for supporting the MEarth Project at FLWO.

**Author Contributions** D.C., Z.K.B., J.I., C.J.B., P.N. and E.E.F. gathered and analysed the photometric data from the MEarth observatory, C.L., X.B., L.A.B., S.U., D.Q., F.P., M.M. and C.J.B. gathered and analysed the spectroscopic data from the HARPS instrument, and L.A.B., D.W.L., M.J.H., J.N.W. and P.N. gathered and analysed supplementary photometric and spectroscopic data with the 1.2-m and 1.5-m FLWO telescopes. R.A.M.-C. estimated the hydrodynamic escape rate, and X.B., X.D., T.F., J.I. and P.N. estimated the properties of the parent star. All authors discussed the results and commented on the manuscript. D.C. led the project and wrote the paper.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.C. (dcharbonneau@cfa.harvard.edu).

LETTERS

# A single sub-kilometre Kuiper belt object from a stellar occultation in archival data

H. E. Schlichting[1,2], E. O. Ofek[1], M. Wenz[3], R. Sari[1,4], A. Gal-Yam[5], M. Livio[6], E. Nelan[6] & S. Zucker[7]

**The Kuiper belt is a remnant of the primordial Solar System. Measurements of its size distribution constrain its accretion and collisional history, and the importance of material strength of Kuiper belt objects[1–4]. Small, sub-kilometre-sized, Kuiper belt objects elude direct detection, but the signature of their occultations of background stars should be detectable[5–9]. Observations at both optical[10] and X-ray[11] wavelengths claim to have detected such occultations, but their implied abundances are inconsistent with each other and far exceed theoretical expectations. Here we report an analysis of archival data that reveals an occultation by a body with an approximately 500-metre radius at a distance of 45 astronomical units. The probability of this event arising from random statistical fluctuations within our data set is about two per cent. Our survey yields a surface density of Kuiper belt objects with radii exceeding 250 metres of $2.1^{+4.8}_{-1.7} \times 10^7$ deg$^{-2}$, ruling out inferred surface densities from previous claimed detections by more than 5$\sigma$. The detection of only one event reveals a deficit of sub-kilometre-sized Kuiper belt objects compared to a population extrapolated from objects with radii exceeding 50 kilometres. This implies that sub-kilometre-sized objects are undergoing collisional erosion, just like debris disks observed around other stars.**

A small Kuiper belt object (KBO) crossing the line of sight to a star will partially obscure the stellar light, an event which can be detected in the star's light curve. For visible light, the characteristic scale of diffraction effects, known as the Fresnel scale, is given by $(\lambda a/2)^{1/2} \approx 1.3$ km, where $a \approx 40$ astronomical units (AU) is the distance to the Kuiper belt and $\lambda \approx 600$ nm is the wavelength of our observations.

Diffraction effects will be apparent in the star's light curve as a result of occulting KBOs provided that both the star and the occulting object are smaller than the Fresnel scale[12,13]. Occultations by objects smaller than the Fresnel scale are in the Fraunhofer regime. In this regime the diffraction pattern is determined by the size of the KBO and its distance to the observer, the angular size of the star, the wavelength range of the observations and the impact parameter between the star and the KBO (see Supplementary Information for details). The duration of the occultation is approximately given by the ratio of the Fresnel scale to the relative velocity perpendicular to the line of sight between the observer and the KBO. Because the relative velocity is usually dominated by the Earth's velocity around the Sun, which is 30 km s$^{-1}$, typical occultations only last a short time of the order of a tenth of a second.

Extensive ground-based efforts have been conducted to look for optical occultations[9,10,14,15]. So far, these visible searches have announced no detections in the region of the Kuiper belt (30–60 AU), but one of these quests claims to have detected some events beyond 100 AU and at about 15 AU (ref. 10). Unfortunately, ground-based surveys may suffer from a high rate of false-positives owing to atmospheric scintillation, and lack the stability of space-based platforms. The ground-breaking idea to search for occultations in archival RXTE X-ray data resulted in several claimed occultation events[11]. Later, revised analysis of the X-ray data[16–19] concluded that the majority of the originally reported events are most probably due to instrumental dead-time effects. Thus, previous reports of optical and X-ray events remain dubious[14] and their inferred KBO abundance is inconsistent with the observed break in the KBO size distribution, which was obtained from direct detections of large KBOs[20–22]. Furthermore, they are also difficult to reconcile with theoretical expectations, which predict collisional evolution for KBOs smaller than a few kilometres in size[4,23] and hence a lower KBO abundance than inferred from extrapolation from KBOs with radii $r > 50$ km.

For the past 14 years, the Fine Guidance Sensors (FGS) on board the Hubble Space Telescope (HST) have been collecting photometric measurements of stars with 40-Hz time resolution, allowing for the detection of the occultation diffraction pattern rather than a simple decrease in the photon count. We examined four and a half years of archival FGS data, which contain about 12, 000 star hours of low ecliptic latitude ($|b| < 20^\circ$) observations.

Our survey is most likely to detect occultations by KBOs that are 200–500 m in radius given the signal-to-noise of our data (Supplementary Fig. 1) and a power-law size distribution with power-law index between 3 and 4.5. Occultation events in this size range are in the Fraunhofer regime, where the diffraction pattern is independent of the occulting object's shape and the depth of the diffraction pattern varies linearly with the area of the object. The theoretical light curves for our search algorithm were therefore calculated in this regime. We fitted these theoretical occultation templates to the FGS data and performed $\chi^2$ analysis to identify occultation candidates (see Supplementary Information). We detected one occultation candidate, at ecliptic latitude 14°, that significantly exceeds our detection criterion (Fig. 1). The best-fit parameters yield a KBO size of $r = 520 \pm 60$ m and a distance of $45^{+5}_{-4}$ AU where we assumed a circular KBO orbit and an inclination of 14°. Using bootstrap simulations, we estimate a probability of ~2% that such an event is caused by statistical fluctuations over the whole analysed FGS data set (Supplementary Fig. 5). We note that for objects on circular orbits around the Sun two solutions can fit the duration of the event. However, the other solution is at a distance of 0.07 AU from the Earth, and is therefore unlikely. It is also unlikely that the occulting object was located in the asteroid belt, because the expected occultation rate from asteroids is about two orders of magnitude less than our implied rate. Furthermore, an asteroid would have to have an eccentricity of order unity to be able to explain the duration of the observed occultation event.

Using the KBO ecliptic latitude distribution from ref. 24, our detection efficiency, and our single detection, we constrain the surface

[1]Department of Astronomy, 249-17, California Institute of Technology, Pasadena, California 91125, USA. [2]CITA, University of Toronto, 60 St George Street, Ontario, M5S 3H8, Canada. [3]Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, Maryland 20771, USA. [4]Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel. [5]Faculty of Physics, Weizmann Institute of Science, POB 26, Rehovot 76100, Israel. [6]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [7]Department of Geophysics and Planetary Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

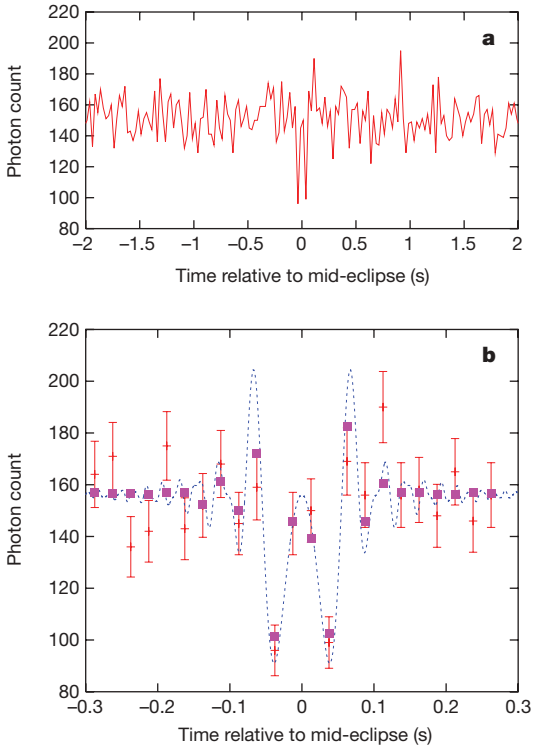



**Figure 1 | Photon counts as a function of time of the candidate occultation event observed by FGS2.** **a**, The photon count spanning ±2 s around the occultation event. **b**, The event in detail. The red crosses are the FGS data points with Poisson error bars, the dashed blue line is the theoretical diffraction pattern (calculated for the 400–700 nm wavelength range of the FGS observations), and the pink squares correspond to the theoretical light curve integrated over 40-Hz intervals. We note that the actual noise for this observation is about 4% larger than Poisson noise owing to additional noise sources such as dark counts (about 3–6 counts in a 40-Hz interval), and jitter caused by the displacement of the guide star (by up to 10 mas) from its null position. The mean signal-to-noise ratio in a 40-Hz interval for the roughly half an hour of observations is ~12. The event occurred at Coordinated Universal Time UTC 05:17:49 on 24 March 2007. The best-fit $\chi^2$ is 20.1 with 21 degrees of freedom. The star has an ecliptic latitude of +14. Its angular radius and effective temperature are about 0.3 of the Fresnel scale and about 4,460 K, respectively. These values were derived by fitting the 2MASS[26] JHK and USNO-B1 BR[27] photometry with a black-body spectrum. The position of the star is RA = 186.87872°, dec. = 12.72469° (J2000) and its estimated V-magnitude is 13.4. The auto-correlation function (excluding lag zero) of the photometric time series of this event is consistent with zero within the statistical uncertainty. Each FGS provides two independent photomultiplier (PMT) readings and we confirmed that the occultation signature is present in both of these independent photon counts. We examined the photon counts of the other guide star that was observed by FGS1 at the time of the occultation and confirmed that the occultation signal is only present in the observations recorded by FGS2. We examined the engineering telemetry for HST around the time of the event and verified that the guiding performance of HST was normal. We therefore conclude that the above occultation pattern is not caused by any known instrumental artefacts.

density around the ecliptic (averaged over $-5^\circ < b < 5^\circ$) of KBOs with radii larger than 250 m to $2.1^{+4.8}_{-1.7} \times 10^7$ deg$^{-2}$ (see Supplementary Information Sections 5 and 6). This surface density is about three times the implied surface density at 5.5° ecliptic latitude and about five times the surface density at 8–20° ecliptic latitude. This is the first measurement of the surface density of hectometre-sized KBOs and it improves previous upper limits by more than an order of magnitude[9,15].

Figure 2 displays our measurement for the sub-kilometre KBO surface density and summarizes published upper limits from various surveys. Our original data analysis focused on the detection of KBOs located at the distance of the Kuiper belt between 30 AU and 60 AU. To compare our results with previously reported ground-based detections

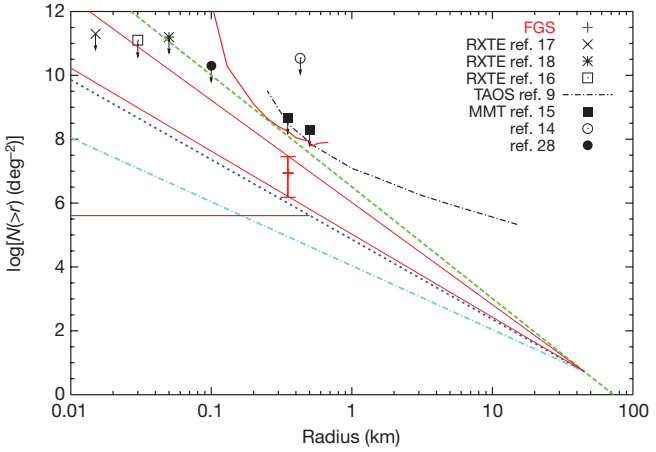



**Figure 2 | Cumulative KBO size distribution as a function of KBO radius for objects located between 30 and 60 AU.** The results from our FGS survey are shown in red and are presented in three different ways. (1) The red cross is derived from our detection and represents the KBO surface density around the ecliptic (averaged over $-5^\circ < b < 5^\circ$) and is shown with 1$\sigma$ error bars. The cross is plotted at $r = 250$ m, which is roughly the peak of our detection probability (see Supplementary Information Section 6 for details). (2) The upper and lower red curves correspond to our upper and lower 95% confidence levels which were derived without assuming any size distribution. (3) The region bounded by the two straight red lines falls within 1$\sigma$ of our best estimate for the power-law size distribution index, that is, $q = 3.9 \pm 0.3$, which was calculated for low ecliptic latitudes ($|b| < 5^\circ$). These lines are anchored to the observed surface density at $r = 45$ km. For comparison, we also show three other lines (green, blue, turquoise). The green (long-dashed) line is the observed size distribution of large KBOs (that is, $r > 45$ km), which has $q = 4.5$, extrapolated as a single power-law to small sizes. The blue (short-dashed) line is a double power-law with $q = 3.5$ (collisional cascade of strength-dominated bodies) for KBOs with radii less than 45 km and $q = 4.5$ above. The turquoise (dot-dashed) line corresponds to $q = 3.0$ (collisional cascade of strengthless rubble piles) for KBOs below 45 km in size. All distributions are normalized to $N(>r) = 5.4$ deg$^{-2}$ at a radius of 45 km (ref. 25). In addition, 95% upper limits from various surveys are shown in black (refs 17, 18, 16, 9, 15, 14 and 28). We note that a power-law index of 3.9 was used for calculating the cumulative KBO number density from the RXTE observations.

beyond 100 AU (ref. 10), we performed a second search of the FGS data that was sensitive to objects located beyond the classical Kuiper belt. Our results challenge the reported ground-based detections of two 300 m-sized objects beyond 100 AU (ref. 10). Given our total number of star hours and a detection efficency of 3% for 300-m-sized objects at ~100 AU we should have detected more than twenty occultations. We therefore rule out the previously claimed optical detections[10] by more than 5$\sigma$. This result accounts for the broad latitude distribution of our observations (that is, $|b| < 20^\circ$) and the quoted detection efficiency of our survey includes the effect of the finite angular radii of the guide stars at 100 AU.

The KBO cumulative size distribution is parameterized by $N(>r) \propto r^{1-q}$, where $N(>r)$ is the number of objects with radii greater than $r$, and $q$ is the power-law index. The power-law index for KBOs with radii above ~45 km is ~4.5 (refs 21, 22) and there is evidence for a break in the size distribution at about $r_{\rm break} \approx 45$ km (refs 20–22). Hence we use this break radius and assume a surface density for KBOs larger than $r_{\rm break}$ (ref. 25) of 5.4 deg$^{-2}$ around the ecliptic. Accounting for our detection efficiency, the velocity distribution of the HST observations, and assuming a single power-law for objects with radii less than 45 km in size, we find $q = 3.9^{+0.3,+0.4}_{-0.3,-0.7}$ (1$\sigma$ and 2$\sigma$ errors) below the break. Our results firmly show a deficit of sub-kilometre-sized KBOs compared to large objects. This confirms the existence of the previously reported break and establishes a shallower size distribution extending two orders of magnitude in size down to sub-kilometre-sized objects. This suggests that sub-kilometre-sized KBOs underwent collisional evolution, eroding the smaller KBOs.

This collisional grinding in the Kuiper belt provides the missing link between large KBOs and dust, producing debris disks around other stars. Currently, our results are consistent with a power-law index of strength-dominated collisional cascade[23], $q = 3.5$, within $1.3\sigma$ and with predictions for strengthless rubble piles[4], $q = 3.0$, within $2.4\sigma$. An intermediate value of $3 < q < 3.5$ implies that KBOs are strengthless rubble piles above some critical size, $r_c < r < 45$ km, and strength-dominated below it, $r < r_c$. Our observations constrain $r_c$ for the first time to our knowledge. At the $2\sigma$ level we find $r_c > 3$ km.

Using our estimate for the size distribution power-law index ($q = 3.9$) and our KBO surface density for 250-m-sized KBOs at an ecliptic latitude of $b = 5.5^\circ$, which is the ecliptic latitude of the RXTE observations of Scorpius X-1, we predict that there should be about $3.6 \times 10^9$ objects of radius 30 m per square degree. This is about 150 times less than the original estimate from X-ray observations of Scorpius X-1 that reported 58 events[11], and it is about 30 times less than the revised estimate from the same X-ray observations, which concludes that up to 12 events might be actual KBO occultations[16]. Our results rule out the implied surface density from these 12 events at $7\sigma$ confidence level. One can reconcile our results and the claimed X-ray detections only by invoking a power-law index of $q \approx 5.5$ between 250 m and 30 m. More recent X-ray work reports no new detections in the region of the Kuiper belt but places an upper limit of $1.7 \times 10^{11}$ deg$^{-2}$ for objects of 50 m in radius and larger[18]. This is consistent with the KBO surface density of $N(>50\,\mathrm{m}) = 8.2 \times 10^8$ deg$^{-2}$ that we derive by extrapolating from our detection in the hectometre size range.

The statistical confidence level on our detection is 98%. However, our conclusions that there is a significant break in the size distribution and that collisional erosion is taking place and the significant discrepancy with previously claimed occultation detections rely on the low number of events we discovered. These conclusions would only be strengthened if this event was caused by an unlikely statistical fluctuation or an as-yet-unknown instrumental artefact.

Ongoing analysis of the remaining FGS data, which will triple the number of star hours, together with further development of our detection algorithm (that is, including a larger number of light-curve templates) holds the promise of additional detections of occultation events and will allow us to constrain the power-law index of the size distribution further.

**Received 12 August; accepted 21 October 2009.**


1. Davis, D. R. & Farinella, P. Collisional evolution of Edgeworth-Kuiper belt objects. *Icarus* **125,** 50–60 (1997).
2. Stern, S. A. & Colwell, J. E. Collisional erosion in the primordial Edgeworth-Kuiper belt and the generation of the 30–50 AU Kuiper gap. *Astrophys. J.* **490,** 879–882 (1997).
3. Kenyon, S. J. & Luu, J. X. Accretion in the early Kuiper belt. II. Fragmentation. *Astron. J.* **118,** 1101–1119 (1999).
4. Pan, M. & Sari, R. Shaping the Kuiper belt size distribution by shattering large but strengthless bodies. *Icarus* **173,** 342–348 (2005).
5. Bailey, M. E. Can 'invisible' bodies be observed in the Solar System? *Nature* **259,** 290–291 (1976).
6. Dyson, F. J. Hunting for comets and planets. *Q. J. R. Astron. Soc.* **33,** 45–57 (1992).
7. Axelrod, T. S., Alcock, C., Cook, K. H. & Park, H.-S. in *Robotic Telescopes in the 1990s* (ed. Filippenko, A. V.) 171–181 (1992).
8. Roques, F., Moncuquet, M. & Sicardy, B. Stellar occultations by small bodies—diffraction effects. *Astron. J.* **93,** 1549–1558 (1987).
9. Zhang, Z.-W. *et al.* First results from the Taiwanese-American Occultation Survey (TAOS). *Astrophys. J.* **685,** L157–L160 (2008).
10. Roques, F. *et al.* Exploration of the Kuiper belt by high-precision photometric stellar occultations: first results. *Astron. J.* **132,** 819–822 (2006).
11. Chang, H.-K. *et al.* Occultation of X-rays from Scorpius X-1 by small trans-neptunian objects. *Nature* **442,** 660–663 (2006).
12. Roques, F. & Moncuquet, M. A detection method for small Kuiper belt objects: the search for stellar occultations. *Icarus* **147,** 530–544 (2000).
13. Nihei, T. C. *et al.* Detectability of occultations of stars by objects in the Kuiper belt and Oort cloud. *Astron. J.* **134,** 1596–1612 (2007).
14. Bickerton, S. J., Kavelaars, J. J. & Welch, D. L. A Search for sub-km Kuiper belt objects with the method of serendipitous stellar occultations. *Astron. J.* **135,** 1039–1049 (2008).
15. Bianco, F. B. *et al.* A Search for occultations of bright stars by small Kuiper belt objects using Megacam on the MMT. *Astron. J.* **138,** 568–578 (2009).
16. Chang, H.-K., Liang, J.-S., Liu, C.-Y. & King, S.-K. Millisecond dips in the RXTE/PCA light curve of Sco X-1 and trans-Neptunian object occultation. *Mon. Not. R. Astron. Soc.* **378,** 1287–1297 (2007).
17. Jones, T. A., Levine, A. M., Morgan, E. H. & Rappaport, S. Production of millisecond dips in Sco X-1 count rates by dead time effects. *Astrophys. J.* **677,** 1241–1247 (2008).
18. Liu, C.-Y., Chang, H.-K., Liang, J.-S. & King, S.-K. Millisecond dip events in the 2007 RXTE/PCA data of Sco X-1 and the trans-Neptunian object size distribution. *Mon. Not. R. Astron. Soc.* **388,** L44–L48 (2008).
19. Blocker, A. W., Protopapas, P. & Alcock, C. R. A Bayesian approach to the analysis of time symmetry in light curves: reconsidering Scorpius X-1 occultations. *Astrophys. J.* **701,** 1742–1752 (2009).
20. Bernstein, G. M. *et al.* The size distribution of trans-neptunian bodies. *Astron. J.* **128,** 1364–1390 (2004).
21. Fuentes, C. I. & Holman, M. J. A SUBARU archival search for faint trans-neptunian objects. *Astron. J.* **136,** 83–97 (2008).
22. Fraser, W. C. *et al.* The Kuiper belt luminosity function from m(R)=21 to 26. *Icarus* **195,** 827–843 (2008).
23. Dohnanyi, J. W. Collisional models of asteroids and their debris. *J. Geophys. Res.* **74,** 2531–2554 (1969).
24. Elliot, J. L. *et al.* The Deep Ecliptic Survey: a search for Kuiper belt objects and centaurs. II. Dynamical classification, the Kuiper belt plane, and the core population. *Astron. J.* **129,** 1117–1162 (2005).
25. Fuentes, C. I., George, M. R. & Holman, M. J. A Subaru pencil-beam search for m(R) ~ 27 trans-neptunian bodies. *Astrophys. J.* **696,** 91–95 (2009).
26. Skrutskie, M. F. *et al.* The Two Micron All Sky Survey (2MASS). *Astron. J.* **131,** 1163–1183 (2006).
27. Monet, D. G. *et al.* The USNO-B catalog. *Astron. J.* **125,** 984–993 (2003).
28. Roques, F., Georgevits, G. & Doressoundiram, A. *The Kuiper Belt Explored by Serendipitous Stellar Occultations* 545–556 (University of Arizona Press, 2008).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank H. K. Chang for comments that helped to improve this manuscript. Some of the numerical calculations presented here were performed on Caltech's Division of Geological and Planetary Sciences Dell cluster. Partial support for this research was provided by NASA through a grant from the Space Telescope Science Institute. R.S. acknowledges support from the ERC and the Packard Foundation. A.G.-Y. is supported by the Israeli Science Foundation, an EU Seventh Framework Programme Marie Curie IRG fellowship and the Benoziyo Center for Astrophysics, a research grant from the Peter and Patricia Gruber Awards, and the William Z. and Eda Bess Novick New Scientists Fund at the Weizmann Institute. S.Z. acknowledges support from the Israel Science Foundation–Adler Foundation for Space Research. E.O.O. is an Einstein Fellow.

**Author Contributions** H.E.S. wrote the detection algorithm, analysed the FGS data for occultation events, calculated the detection efficiency of the survey, preformed the bootstrap analysis and wrote the paper. E.O.O. calculated the stellar angular radii, the velocity information of the observations, the correlated noise and other statistical properties of the data. R.S. guided this work and helped with the scientific interpretation of the results. A.G.-Y. proposed using HST FGS data for occultation studies and helped to make the data available for analysis. M.W. extracted the FGS photometry streams and provided coordinates and magnitudes of the guide stars. M.L. helped in gaining access to the FGS data and provided insights into the operation and noise properties of the FGS. E.N. provided expert interpretation of the FGS photometric characteristics in the HST operational environment. S.Z. took part in the statistical analysis of the data. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.E.S. (hes@astro.caltech.edu) or E.O.O. (eran@astro.caltech.edu).

Vol 462|17 December 2009|doi:10.1038/nature08563

LETTERS

# Photon-by-photon feedback control of a single-atom trajectory

A. Kubanek[1], M. Koch[1], C. Sames[1], A. Ourjoumtsev[1], P. W. H. Pinkse[1], K. Murr[1] & G. Rempe[1]

**Feedback is one of the most powerful techniques for the control of classical systems. An extension into the quantum domain is desirable as it could allow the production of non-trivial quantum states[1–4] and protection against decoherence[5,6]. The difficulties associated with quantum, as opposed to classical, feedback arise from the quantum measurement process—in particular the quantum projection noise and the limited measurement rate—as well as from quantum fluctuations perturbing the evolution in a driven open system. Here we demonstrate real-time feedback control[7–12] of the motion of a single atom trapped in an optical cavity. Individual probe photons carrying information about the atomic position[13,14] activate a dipole laser that steers the atom on timescales 70 times shorter than the atom's oscillation period in the trap. Depending on the specific implementation, the trapping time is increased by a factor of more than four owing to feedback cooling, which can remove almost all the kinetic energy of the atom in a quarter of an oscillation period[12]. Our results show that the detected photon flux reflects the atomic motion, and thus mark a step towards the exploration of the quantum trajectory[15,16] of a single atom at the standard quantum limit.**

In contrast to highly energetic charged particles, whose trajectories can be observed using ionization detectors, single neutral atoms are much harder to track. The reason is that the interaction of neutral particles with a detector is much weaker than that of charged particles. So far, the most efficient means of detecting single atoms is light scattering. However, the scattering rate is limited by the natural decay rate of the atom's excited state and the photons are emitted in random directions. Therefore, the signal becomes vanishingly small if rapid measurements must be performed, for example in the implementation of fast feedback on single atoms perturbed by quickly changing random forces.

In this context, optical cavity quantum electrodynamics provides a powerful technique for single-atom tracking and feedback owing to its unique observation[13,14] and control capabilities[9,17], respectively. Measurements are faster and more sensitive than in free space, as a result of the increased rate of information exchange between the atom and the observed cavity field. This makes it possible to estimate the atomic trajectory quickly and use this position information to steer the atom rapidly in the desired direction. An advantage of such a strategy is that the steering force is automatically synchronized with the atomic motion. This is useful if the atom is moving in an anharmonic potential (as here) where the oscillation period depends on the oscillation amplitude, in which case an actuator operated at a fixed frequency is insufficient. More importantly, it makes it possible to control the atomic motion even if this motion is unpredictable on timescales as short as the oscillation period in the trap.

Unlike in previous work[9,10], the actuator uses blue-detuned dipole light that pushes the atom towards the area of low light intensity in the centre of the cavity[17]. This has three benefits, which we found to be essential. First, the dipole laser induces hardly any shift in the energy levels of the atom, so tracking and steering are largely independent of each other. Second, the dipole laser controls the motion perpendicular to the cavity axis, which is the typical escape direction for a trapped atom. Third, it leaves the atomic motion along the cavity axis unperturbed and does not interfere with cavity cooling along this direction. As a result, we are able to study the deterministic, as well as the probabilistic, nature of the atomic trajectory by tuning the observation time interval and measuring the system response.

The system is sketched in Fig. 1. A high-finesse cavity supports a $TEM_{00}$ mode (waist, ~29 μm) nearly resonant with the transition

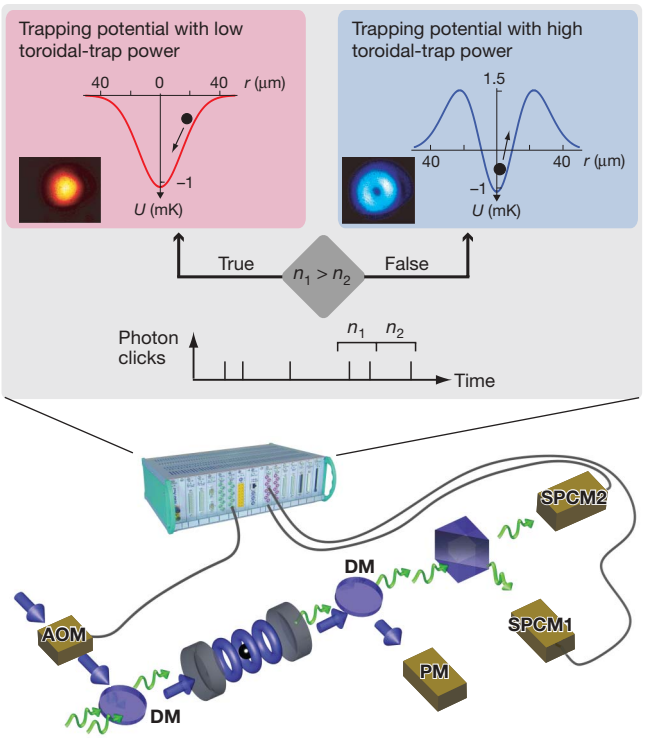



**Figure 1 | Experimental setup including the feedback loop.** An optical cavity directs the transmitted light to two single-photon counting modules (SPCM1, SPCM2). A real-time processor determines the sums of the photon clicks of both detectors in two consecutive time windows, $n_1$ and $n_2$, of equal duration, $T$. To reduce the kinetic energy of the atom, the algorithm switches the toroidal blue-detuned dipole trap to high power if the atom attempts to leave the trap (here for $n_1 \leq n_2$) and to a low power if the atom returns towards the cavity axis ($n_1 > n_2$). The power of the toroidal trap is switched using an acousto-optical modulator (AOM), superimposed with the probe light using a dichroic mirror (DM) and detected using a photomultiplier (PM). The resulting trapping potential, $U$, is plotted as a function of the radial distance from the cavity centre, $r$, for low power (red box) and high power (blue box).

[1]Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Strasse 1, D-85748 Garching, Germany.

from the $5^2S_{1/2}$, $F=3$, $m_F=3$ state to the $5^2P_{3/2}$, $F=4$, $m_F=4$ state in $^{85}$Rb atoms (wavelength, 780 nm), which results in a maximum atom–cavity coupling of $g_0/2\pi = 16$ MHz, exceeding losses due to atomic polarization decay (rate, $\gamma/2\pi = 3$ MHz) and cavity field decay (rate, $\kappa/2\pi = 1.25$ MHz). A weak probe laser excites this mode at the input mirror. In addition, a red-detuned dipole-trap laser of wavelength 785 nm (not drawn) resonantly excites a second $TEM_{00}$ mode[18], confining the atom in the cavity (trap depth, ~1 mK). A separate, blue-detuned, dipole trap is implemented at 775 nm and consists of a $TEM_{10}$ and a $TEM_{01}$ mode, each of which is blue-detuned with respect to the probe light by two free spectral ranges. Together they form a toroidal repulsive trap[17]. Because in a cylindrically symmetric system circular orbits do not modulate the transmitted light, and because cylindrically symmetric forces cannot change the angular momentum of the atomic trajectory, we break the cylindrical symmetry by making the $TEM_{01}$ component 50% stronger than the $TEM_{10}$ component. The light exiting through the output mirror is separated into probe light and trap light, and the probe light is split by a non-polarizing beam splitter and detected using two single-photon counting modules. The probe laser is almost resonant with the empty cavity (detuned by $2\pi \times 100$ kHz) and is detuned from the Stark-shifted atomic resonance by $2\pi \times 20$ MHz. For such parameter choices, a well-coupled atom will induce a drop in the transmission from 1 photon per microsecond (for an empty cavity) to typical values as low as 0.03 photons per microsecond.

Our digital feedback algorithm uses the blue toroidal trap as an 'actuator'. The input signal is sensitive to the atomic trajectory in real time. An increase in the transmission indicates a lower coupling to the mode, as happens when an atom leaves the cavity. This information is extracted by the feedback processor (ADwin-Pro II system), which compares the respective numbers of photon clicks, $n_1$ and $n_2$, registered during two consecutive user-defined intervals of equal duration, $T$, the exposure time. Differential feedback routines are used to switch the torus potential whenever a turning point of the atomic trajectory in the radial direction is registered. To achieve a high efficiency, we apply a 'bang-bang' strategy[12], in which the intensity of the torus potential is switched, using an acousto-optical modulator, between two extreme powers: 50 nW (low), resulting in an overall trap depth of 1 mK, and 800 nW (high), corresponding to 2.5 mK.

We implement two feedback strategies, which work as follows. In the 'normal' feedback strategy, we decrease the kinetic energy of the atom and keep it in the cavity centre. This is done by switching the toroidal trap to high power as soon as the atom attempts to leave the cavity, and switching it to low power as soon as the atom moves towards the cavity axis; see Fig. 1 and Fig. 2a. In the 'inverted' feedback strategy, the switching protocol is reversed, increasing the kinetic energy of the atom and expelling it from the cavity.

To be able to follow these strategies as closely as possible, we designed fast digital feedback logic that can react to the detection of a single photon with a decision-making time of 1.7 μs and a maximum switching-process delay of 3 μs. This is much faster than the radial oscillation, which has a period of ~360 μs, so feedback occurs in real time. We note that the transmission signal is modulated at half the oscillation period, $T_{osc}$, of the atom, owing to the symmetry of the cavity modes. For exposure times $T \approx T_{osc}$, the signal becomes averaged, increasing the signal-to-noise ratio but delaying the algorithm. For $T < T_{osc}$, the algorithm is fast but only partial information about the atomic trajectory is acquired. We assume that the atom leaves the cavity radially for an increasing photon flux ($n_1 < n_2$) and that the atom returns towards the cavity axis for a decreasing photon flux ($n_1 > n_2$). Both feedback strategies are shown in Fig. 2b for a short exposure time, $T = 10.2$ μs. Here the $n_1$ and $n_2$ values are mostly zero. Nevertheless, the normal feedback strategy keeps the photon rate low, indicating good confinement of the atom. In contrast, the inverted feedback strategy leads to an increasing photon flux. Figure 2c shows an example for a longer exposure time, $T = 85$ μs. Here the integrated signal becomes a smooth function, resulting in smoother switching.

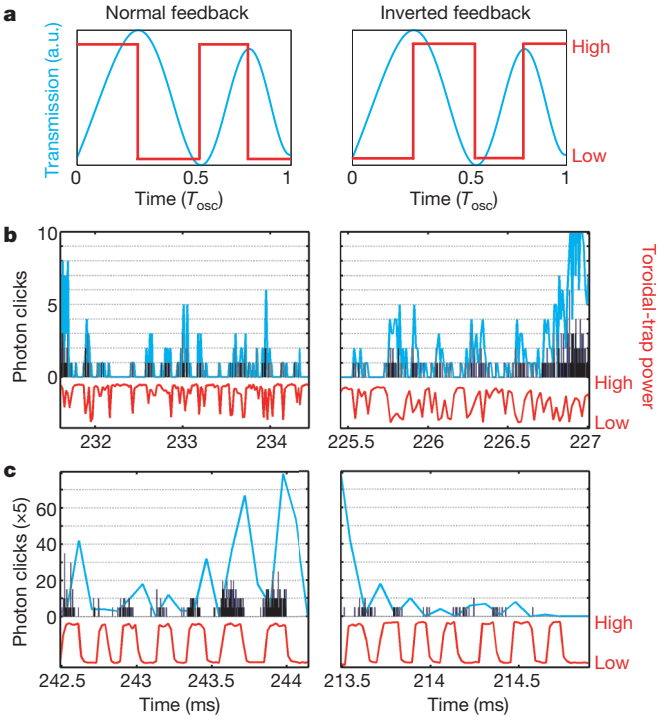



**Figure 2 | Feedback protocol for a single-atom trajectory.** **a**, The idealized transmission (blue) and feedback reaction (red) are displayed for an atom oscillating in the trap. For normal feedback (left column), the toroidal trap is switched to low power for decreasing cavity transmission and to high power otherwise. For inverted feedback (right column), the switching behaviour is reversed. a.u. arbitrary units. **b**, The two strategies are shown for a short exposure time, $T = 10.2$ μs. Single photon clicks are indicated in black and their integration over the exposure time is indicated in blue. Normal feedback keeps the atom in a well-coupled regime, reducing the oscillation amplitude, whereas inverted feedback is out of phase with the atom's oscillation and therefore increases the oscillation amplitude to the point at which the atom leaves the trap. **c**, For a long exposure time, $T = 85$ μs, inverted feedback is in phase, reducing the oscillation, whereas normal feedback is not. Data for single photon clicks is shown multiplied by a factor of five, for clarity.

Concurrently, a large exposure time also delays the feedback, which can be seen by comparing the raw photon clicks (black data) with the integrated signal (blue data). For this exposure time, normal feedback is out of phase with the radial oscillation and therefore drives the atom out of the cavity. As inverted feedback is also delayed by one-quarter of an oscillation period, it turns into a damping force.

Special attention must be paid to the case in which $n_1 = n_2$. We found that the feedback is effective only if we switch the toroidal trap to low power for $n_1 > n_2$ and to high power for $n_1 \leq n_2$. A possible explanation is that for short exposure times, $T$, the zero-photon events are the most likely for atoms near the cavity axis. For the chosen, near-resonant, probing of the system, the photon flux depends only marginally on the exact atomic position when the coupling strength is large. Hence, if $n_1 = n_2 = 0$ the atom could already be moving away from the axis, in which case it is prudent to switch the toroidal trap to high power.

To analyse the performance of our feedback strategies quantitatively, we study the atomic storage time, which is mainly determined by radial losses of the atom for the chosen parameters. The storage time is obtained by fitting an exponential decay to the atomic storage probability. The first millisecond is dominated by a rapid loss of atoms, presumably those that are not injected in one of the central antinodes. Therefore, the first millisecond of data is disregarded (fraction of remaining atoms, ~55%).

Figure 3 shows the atomic storage probability as a function of time (plotted in the inset) and the average storage time as a function of

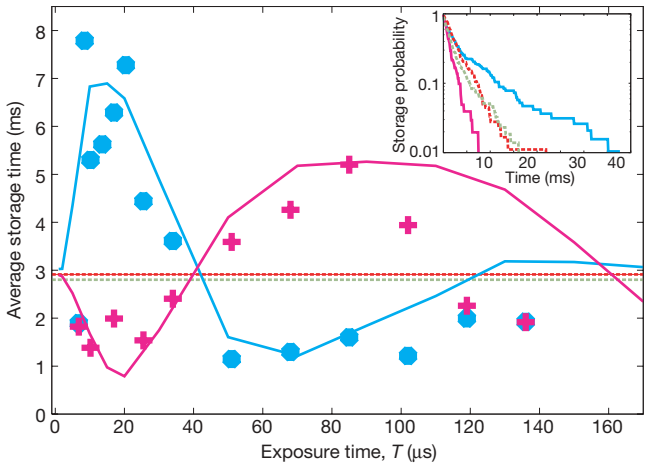



**Figure 3 | Manoeuvring of a single atom using real-time feedback.** Average storage time plotted as a function of exposure time for two feedback strategies. The solid lines show the results of Monte Carlo simulations for the inverted (magenta) and normal (blue) feedback strategies. The corresponding experimental data points are plotted in the same colours. The dotted green and red lines show the experimental storage times without feedback for high and, respectively, low toroidal-trap power, setting the boundary between feedback cooling (longer storage time) and feedback heating (shorter storage time). The standard deviations (not shown) obtained from exponential fits to the decay (see text) are less than 0.18 ms. The inset shows the experimental atom-loss dynamics without feedback (green and red dashed lines, colour-coded as in main figure) and illustrates how the feedback changes the slope of the exponential decay. Whereas the slope is less steep for normal feedback ($T = 17$ μs; blue), it is more steep for inverted feedback ($T = 25.5$ μs; magenta).

exposure time, $T$, ranging from a few microseconds to $T \approx 140$ μs or, in other words, from a small fraction of the oscillation period to about three-eighths of it. The dotted red and green lines are the storage times with the toroidal trap at low and, respectively, high power without feedback. When $T$ is less than ~7 μs, the noise dominates the signal and the algorithm mainly reacts to the random arrival of photons. This results in a heating of the atom and a decrease in the storage time to below the zero-feedback value, for both normal (blue points) and the inverted (magenta points) feedback. Increasing the exposure time, however, increases the storage time, with a maximum at $T \approx 15$ μs for normal feedback. This is remarkable, given that only a few photons have been detected. In contrast, applying inverted feedback at this value of $T$ leads to a storage time less than the value obtained without feedback. Here we are heating the atom by acting out of phase with its motion. Furthermore, we see that inverted feedback turns into a cooling mechanism for longer exposure times, reaching a maximum at $T \approx 90$ μs. The maximum storage time for inverted feedback is less than that obtained for normal feedback. This directly shows that it is important to apply the feedback before the atomic motion becomes unpredictable.

To understand the data better, we performed Monte Carlo simulations of the experiment. As shown by the solid lines in Fig. 3, there is good agreement between the simulations and the measurements. The storage times are well reproduced if we add a loss mechanism other than cavity-induced heating. The main reason for this loss is the possibility of off-resonant pumping of the atom into the 'dark' hyperfine state, in which $F = 2$ (ref. 19). We analysed this effect by adding a repumper laser perpendicular to the cavity axis. From the increase in the storage time without feedback, we can estimate the additional loss rate to be between $1/5.5\,\mathrm{ms}^{-1}$ and $1/14\,\mathrm{ms}^{-1}$. Including a rate of $1/7\,\mathrm{ms}^{-1}$ in the simulation shows good agreement with the experimental data. We also found experimentally that the storage time with feedback increases to ~15 ms when the repumper laser is added.

We next analyse the atomic motion by imposing limits on the average transmitted power (and, thus, on the average atom–cavity coupling strength) in the data evaluation[20]. Photon correlations now reveal further information on the dynamics of atomic motion, as shown in Fig. 4. Here photon bunching is caused by fluctuations in the coupling strength. For measurements with a weak atom–cavity coupling, corresponding to a transmission of up to 0.6 times that of the empty cavity, the oscillation period is ~2 × 260 μs (dashed line). If the transmission is restricted to be less than 0.05 times that of the empty cavity, the oscillation period decreases to ~2 × 180 μs (solid line). Comparison with Fig. 3 therefore shows that the maximum in the average storage time at ~90 μs corresponds to one-quarter of the oscillation period for a well-coupled atom. A further comparison with simulations shows that the feedback can keep the atom close to the cavity axis, with an average excursion of less than 4.5 μm. This is a factor of ~2 less than the value obtained without feedback. We note that the crossing from in-phase behaviour to out-of-phase behaviour occurs for $T \approx 45$ μs (Fig. 3), which is one-eighth of the oscillation period for a well-coupled atom. In addition, the random character of the atomic motion is visible in the damping of the correlation function, indicating a decoherence time of ~200 μs. The dependence of the oscillation period on the localization is shown in the inset of Fig. 4. It shows the anharmonicity of the atomic motion; a harmonic potential would result in a horizontal line.

Further improvement of the feedback algorithm is possible and has already been realized by combining a fast reaction time with smooth switching behaviour: a fast loop with a short exposure time, $T_{\mathrm{short}} = 8.5$ μs, is responsible for switching the toroidal trap to low power whenever the atom is observed to move towards the cavity axis ($n_1 > n_2$). Proper switching of the trap back to high power when the atom is close to the cavity axis is more critical, as here the photon flux from the cavity is low. To improve the signal-to-noise ratio, we added a loop with a longer exposure time, $T_{\mathrm{long}} = 34$ μs, corresponding to about $T_{\mathrm{osc}}/8$. The trap is switched back to high power only if both loops register $n_1 \leq n_2$. Using this improved algorithm, we increased the average storage time to about 24 ms, with maximum observed trapping times exceeding 250 ms. With the repumper laser on, the overall improvement factor of the storage time with feedback (24 ms) relative to that without feedback (≤6 ms) is larger than four. Our simulations indicate that storage times of 0.5 s can be realized when spurious photon clicks (in our experiment mainly stemming from repumper light scattered from the edges of the cavity mirrors) are

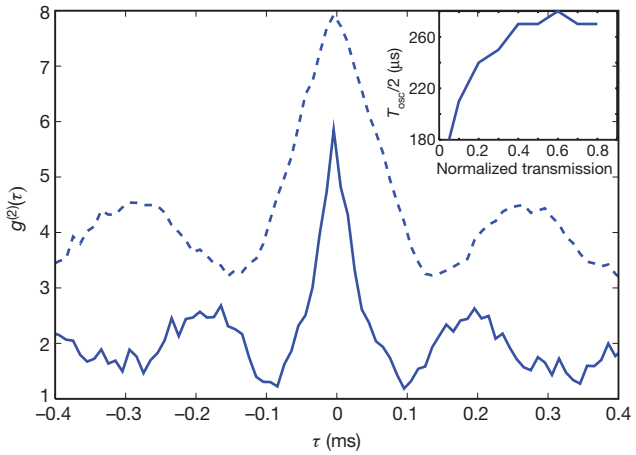



**Figure 4 | Dynamics of atomic motion from photon correlation measurements.** The photon correlation function, $g^{(2)}(\tau)$, is shown as function of the correlation time, $\tau$, for strongly coupled atoms (solid line) and loosely coupled atoms (dashed line). The atomic motion leaves its signature in the correlations, allowing the determination of the oscillation period as well as the decoherence time of the oscillation, ~200 μs. The standard deviations (not shown) of the $g^{(2)}(\tau)$ data are typically less than 0.15. The inset shows the dependence of the oscillation period on the transmission normalized using the empty-cavity transmission, characterizing the anharmonicity of the trapping potential.

eliminated. This would make the feedback scheme compatible with state-of-the-art laser cooling techniques, but with the advantage that one-dimensional optical access is sufficient for three-dimensional control.

The successful realization of feedback on an a-priori unpredictable atomic trajectory shows that reliable position information can be obtained from continuous (or quasi-continuous) measurements. Once extended into the quantum domain, this might make it possible to stabilize the quantum state of a trapped particle or observe the quantum Zeno effect for a free particle[21]. Additional feedback might then make it possible to steer an individual quantum trajectory with a precision ultimately determined by Heisenberg's uncertainty relation.

**Received 17 August; accepted 7 October 2009.**

1. Shapiro, J. H., Saplakoglu, G., Ho, S.-T., Kumar, P. & Saleh, B. E. A. Theory of light detection in the presence of feedback. *J. Opt. Soc. Am. B* **4,** 1604–1620 (1987).
2. Wiseman, H. M. Quantum theory of continuous feedback. *Phys. Rev. A* **49,** 2133–2150 (1994).
3. Jacobs, K. How to project qubits faster using quantum feedback. *Phys. Rev. A* **67,** 030301 (2003).
4. Combes, J., Wiseman, H. M. & Jacobs, K. Rapid measurement of quantum systems using feedback control. *Phys. Rev. Lett.* **100,** 160503 (2008).
5. Viola, L., Knill, E. & Lloyd, S. Dynamical decoupling of open quantum systems. *Phys. Rev. Lett.* **82,** 2417–2421 (1999).
6. Viola, L. Advances in decoherence control. *J. Mod. Opt.* **51,** 2357–2367 (2004).
7. Ashkin, A. & Dziedzic, J. M. Feedback stabilization of optically levitated particles. *Appl. Phys. Lett.* **30,** 202–204 (1977).
8. Morrow, N. V., Dutta, S. K. & Raithel, G. Feedback control of atomic motion in an optical lattice. *Phys. Rev. Lett.* **88,** 093003 (2002).
9. Fischer, T., Maunz, P., Pinkse, P. W. H., Puppe, T. & Rempe, G. Feedback on the motion of a single atom in an optical cavity. *Phys. Rev. Lett.* **88,** 163002 (2002).
10. Lynn, T. W., Birnbaum, K. & Kimble, H. J. Strategies for real-time position control of a single atom in cavity QED. *J. Opt. B* **7,** 215–225 (2005).
11. Bushev, P. *et al.* Feedback cooling of a single trapped ion. *Phys. Rev. Lett.* **96,** 043003 (2006).
12. Steck, D. A., Jacobs, K., Mabuchi, H., Habib, S. & Bhattacharya, T. Feedback cooling of atomic motion in cavity QED. *Phys. Rev. A* **74,** 012322 (2006).
13. Pinkse, P. W. H., Fischer, T., Maunz, P. & Rempe, G. Trapping an atom with single photons. *Nature* **404,** 365–368 (2000).
14. Hood, C. J., Lynn, T. W., Doherty, A. C., Parkins, A. S. & Kimble, H. J. The atom–cavity microscope: single atoms bound in orbit by single photons. *Science* **287,** 1447–1453 (2000).
15. Molmer, K., Castin, Y. & Dalibard, J. Monte Carlo wave-function method in quantum optics. *J. Opt. Soc. Am. B* **10,** 524–538 (1993).
16. Carmichael, H. (ed.) *An Open Systems Approach to Quantum Optics* (Springer, 1993).
17. Puppe, T. *et al.* Trapping and observing single atoms in a blue-detuned intracavity dipole trap. *Phys. Rev. Lett.* **99,** 013002 (2007).
18. Maunz, P. *et al.* Cavity cooling of a single atom. *Nature* **428,** 50–52 (2004).
19. Khudaverdyan, M. *et al.* Quantum jumps and spin dynamics of interacting atoms in a strongly coupled atom–cavity system. *Phys. Rev. Lett.* **103,** 123006 (2009).
20. Kubanek, A. *et al.* Two-photon gateway in one-atom cavity quantum electrodynamics. *Phys. Rev. Lett.* **101,** 203602 (2008).
21. Braginsky, V. B. & Khalili, F. Y. (eds) *Quantum Measurement* (Cambridge Univ. Press, 1992).

**Acknowledgements** Partial support by the Bavarian PhD programme of excellence QCCC, the Deutsche Forschungsgemeinschaft research unit 635 and the European Union project SCALA are gratefully acknowledged.

**Author Contributions** All authors contributed to the design and implementation of the experiment, the interpretation of the results and the writing of the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.K. (alexander.kubanek@mpq.mpg.de) or G.R. (gerhard.rempe@mpq.mpg.de).

Vol 462|17 December 2009|doi:10.1038/nature08662

LETTERS

# Photon-induced near-field electron microscopy

Brett Barwick[1], David J. Flannigan[1] & Ahmed H. Zewail[1]

**In materials science and biology, optical near-field microscopies enable spatial resolutions beyond the diffraction limit[1,2], but they cannot provide the atomic-scale imaging capabilities of electron microscopy[3]. Given the nature of interactions[4–8] between electrons and photons, and considering their connections[9,10] through nanostructures, it should be possible to achieve imaging of evanescent electromagnetic fields with electron pulses when such fields are resolved in both space (nanometre and below) and time (femtosecond)[11–13]. Here we report the development of photon-induced near-field electron microscopy (PINEM), and the associated phenomena. We show that the precise spatiotemporal overlap of femtosecond single-electron packets with intense optical pulses at a nanostructure (individual carbon nanotube or silver nanowire in this instance) results in the direct absorption of integer multiples of photon quanta ($n\hbar\omega$) by the relativistic electrons accelerated to 200 keV. By energy-filtering only those electrons resulting from this absorption, it is possible to image directly in space the near-field electric field distribution, obtain the temporal behaviour of the field on the femtosecond timescale, and map its spatial polarization dependence. We believe that the observation of the photon-induced near-field effect in ultrafast electron microscopy demonstrates the potential for many applications, including those of direct space-time imaging of localized fields at interfaces and visualization of phenomena related to photonics, plasmonics and nanostructures.**

Imaging in conventional electron microscopes is based on elastic interactions of electrons with matter; that is, with no energy loss or gain. With variant techniques, these scattering processes at different angles provide structural and bonding information from images and diffraction patterns[3,14]. In contrast, the interaction is inelastic in electron spectroscopy and what is observed normally is the energy loss due to the valence/core electron excitations of matter; as such, electron energy loss spectroscopy (EELS) is a powerful analytical tool. When images, diffraction, or electron spectra are time-resolved in electron microscopy, photons are typically used to initiate a change for the study of ultrafast structural dynamics, which occur on the femtosecond and longer timescale[11–13]. But, before these structural changes, electronic distributions are altered, with their dynamical changes being on the femtosecond and shorter timescale; they are directly the result of photon–matter interaction. Photon–electron interactions are of a different nature.

In free space, an electron cannot absorb a quantum of electromagnetic energy because of the lack of energy-momentum conservation. However, as suggested in 1933[4], absorption followed by stimulated emission can occur when two (counter-propagating) photons are used. Indeed, the experimental verification was made in 2001[5], and earlier in 1988 it was demonstrated[6] that an intense standing optical wave can result in momentum transfer to free electrons with scattering rates approaching that of the optical frequency. These photon–electron interactions are basic to attosecond pulse generation[7] and to multiphoton harmonics and the laser-assisted surface photoelectric effect[8]. In 1966, the effect of lattice vibrations in electron energy loss and gain was reported[15], and more recently it was suggested[9] that photon–electron interactions can similarly be exploited in various connections. Of special interest is the possible use of continuous-wave (CW, no time resolution) lasers to provide high resolution spectra of plasmons and other features[10]. As shown below, only when the field is induced and probed on the ultrashort timescale would it be visualized and controlled for applications in imaging and spectroscopy.

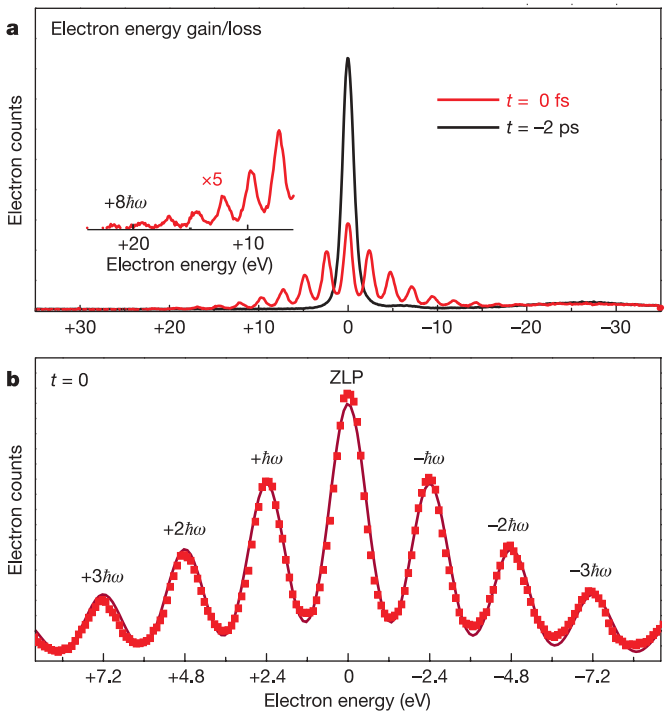



**Figure 1 | Electron energy spectra of carbon nanotubes irradiated with an intense fs laser pulse at two different delay times. a**, The zero-loss peak (ZLP) of the 200 keV electrons (black curve) taken when the electron packet arrives before the femtosecond pulse; in this spectrum only the plasmon peaks are present (see text). The energy spectrum at coincidence of the two pulses ($t = 0$ fs; red curve) displays the multiple quanta of photon absorption/emission. Inset, the positive energy gain region multiplied by 5 for the $t = 0$ spectrum, indicating that absorption of at least eight quanta of photon energy can be observed at maximum spatiotemporal overlap. **b**, Magnified view of the electron energy spectrum obtained at $t = 0$. The energy is given in reference to the loss/gain of photon quanta by the electrons with respect to the zero-loss energy.

Unlike all previous ultrafast electron microscopy (UEM) studies from this laboratory[13], the imaging and spectroscopy methods reported here rely on the strong interaction between photons and electrons. Furthermore, because structural dynamics commence after

[1]Physical Biology Center for Ultrafast Science and Technology, Arthur Amos Noyes Laboratory of Chemical Physics, California Institute of Technology, Pasadena, California 91125, USA.

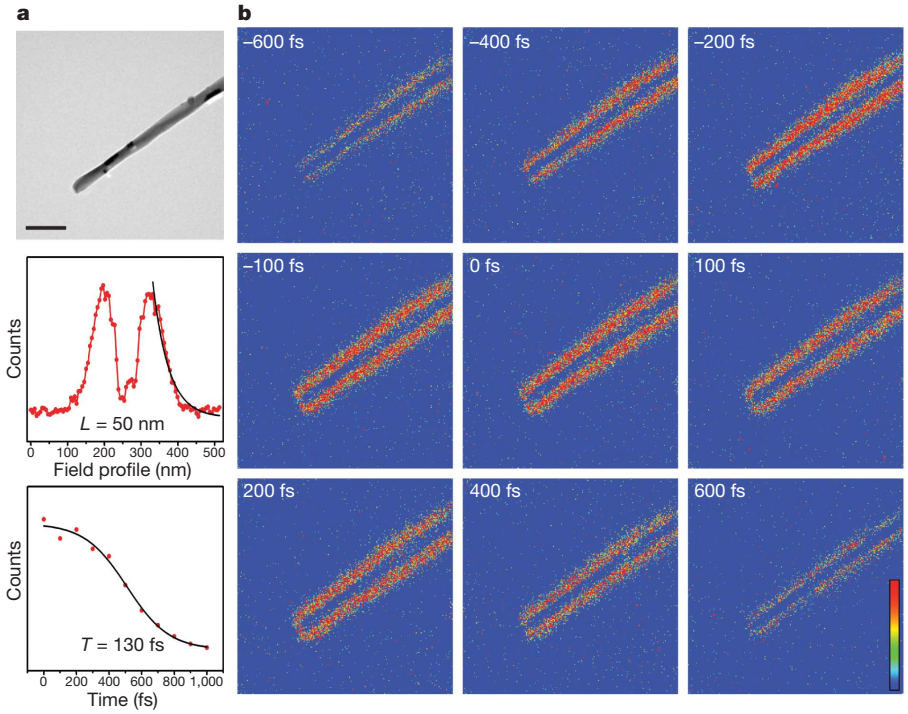



**Figure 2 | Photon-induced near-field electron microscopy of an individual nanotube. a**, Top, bright-field image shown for reference (time-averaged, unfiltered); the average diameter across the tube is 147 ± 20 nm; scale bar, 500 nm. Middle, the spatial field gradient (of length scale $L$) in image counts. Bottom, the decay of counts with time $T$. See text for details. The profile in red shown in the middle plot was obtained from the $t = 0$ frame, and the black exponential curve is displayed simply to illustrate the typical length scale of an evanescent field. **b**, The nine energy-filtered UEM images acquired by using only the electrons that have gained energy (up to $n = 4$) relative to the ZLP; for clarity, the images are displayed in false colour (see bar at lower right). Blue indicates regions of the CCD where no counts were recorded because we selected only the $+n\hbar\omega$ region. The time of arrival of the electron packet at the nanotube relative to the clocking laser pulse is shown in the upper left corner of each image. The electric field of the clocking laser pulse was linearly polarized perpendicular to the long-axis of the nanotube (compare Fig. 3). The counts indicated in red represent the fields created by the femtosecond pulse around the surface of the nanostructure and their decay with time.

the photon-induced near-field (PIN) effect diminishes, here after 400 fs, imaging of the nanostructure electronic properties beyond this time is no longer possible. Significantly, on this electron–photon interaction timescale, and using intense pulses with peak irradiances of the order of 100 GW cm$^{-2}$, the 200 keV electron packets lose and gain energy in discrete quanta that are integer multiples of the tuned photon energy. As reported below, at least eight photons of absorption/emission were observed despite the fact that the interaction time with the nanostructure is only a few hundred attoseconds, given the electron speed and path length in the nanostructure. Moreover, by energy-filtering the zero-loss peak (ZLP), we are able to obtain the images formed as a result of elastic scatterings; remarkably, when selecting only the electrons that have gained quanta of photon energy, the evanescent electric field was visualized in real space images of the nanostructure. The field polarization and temporal behaviour are different from those of bulk structural transformations.

The experiments were performed on an individual and a collection of multiwalled carbon nanotubes with diameters of ~140 nm and lengths of ~7 μm (Aldrich, >90% purity), and on silver nanowires with diameters of ~100 nm and lengths ranging from 2 to 20 μm (ref. 16). Both femtosecond-resolved electron energy spectra and energy-filtered photon-induced images were recorded in our second generation, ultrafast electron microscope (UEM-2). All experiments reported here were conducted in the single-electron regime (0.1–1 e$^-$ per packet at the detector) in order to eliminate space-charge effects; a repetition rate of 500 kHz and a fluence of 14 mJ cm$^{-2}$ for carbon nanotubes (1.2 mJ cm$^{-2}$ for silver nanowires) were typically used. The output from a laser emitting a train of 220 fs pulses centred at 1,038 nm was split into two arms, one of which was frequency doubled and used to excite the nanostructure, and the other was frequency tripled and used to generate the electron packets at the photocathode source in UEM-2. The femtosecond timing between the frequency doubled laser pulse and the electron packet at the specimen was controlled by an optical delay line; the apparatus has been described in detail elsewhere[11,17].

Shown in Fig. 1 are temporally resolved electron energy spectra obtained from the carbon nanotubes irradiated with the intense femtosecond laser pulse. For comparison, two separate electron spectra are shown (Fig. 1a): one obtained when the electron packet arrived before the femtosecond pulse, which we call the $t = -2$ ps spectrum, and a second one recorded when the electron and photon pulses were configured for a maximum overlap, the $t = 0$ fs spectrum. Upon inspection of the change, one immediately notices that at the maximum overlap ($t = 0$) the spectrum consists of discrete peaks of decreasing intensity on both the lower- and higher-energy side of the ZLP. For the $t = -2$ ps spectrum, however, the discrete peaks are absent and only the $\pi$ and $\pi + \sigma$ plasmon peaks at 6 and 25 eV, respectively, are observed[18].

A magnified view of the electron spectrum obtained at $t = 0$ reveals that the discrete peaks on both sides of the ZLP occur at integer multiples of the photon energy of the exciting femtosecond pulse (2.4 eV; Fig. 1b). From these spectra, it is apparent that the discrete peaks occur as a consequence of the interaction of the 200 keV ultrafast electron packet with the 2.4 eV femtosecond photon pulse. On this timescale, the large influence of the PIN effect is illustrated by the substantial decrease in the ZLP intensity at maximum overlap. From the recorded electron-energy spectra, we found that electrons of the ultrafast packets can absorb more than eight photons during the brief interaction with the nanostructure (Fig. 1a inset). It is important to note that the spectra shown in Fig. 1 (loss/gain) are observed only in the presence of the nanostructure.

On the length scale ($d$) of the nanostructure, relative to that of the photon (wavelength $\lambda$), the interaction between photons and electrons (that is, free–free transitions) is greatly enhanced by the evanescent field that is created through the excitations of the carbon

nanotube, and similarly for the silver nanowires, without which energy-momentum conservation is impossible[10,19]. The probability of electron–photon coupling in the presence of a third body (for example, atom, molecule, or a surface) increases as the electron energy increases for a fixed laser intensity and wavelength[8], and such a characteristic is ideal for UEM at 200 keV. Free–free transitions in the electron continuum, without perturbations from the third body, become descriptive of the process when the electron has a much higher energy than the photon[8,20,21]. Whereas plasmonic fields induced by the femtosecond pulse can follow the laser electric field[22], the relaxation time for metallic nanoparticles is of the order of tens to hundreds of femtoseconds, depending on the damping (recombination) processes involved[23]. Multiwalled carbon nanotubes are metallic in nature[24], as are silver nanowires, and for such metallic nanostructures absorption is enhanced at lower energies[2].

In Fig. 2, the effect of energy-filtering on imaging is displayed at different times. By using an energy filter ($n = 1$ to 4; 10 eV total width) to select only those electrons that have absorbed energy to form an image, the evanescent field generated by the femtosecond excitation pulse became evident in real-space images of the isolated (individual) nanotube. Further, by varying the arrival time of the electron packet at the carbon nanotube relative to the clocking femtosecond pulse, the ultrafast evolution of the evanescent field was followed in real time. As can be seen in Fig. 2, image counts appear only within the local vicinity of the surface of the carbon nanotube; no energy gain occurs far from the nanostructure or within the tube itself.

The energy-filtered image generated by selecting only energy-gained electrons and obtained at $t = -600$ fs shows almost no counts. As the temporal overlap increases, however, the counts due to the evanescent optical field increases and reaches a maximum at $t = 0$ (that is, maximum overlap) before decreasing again to almost zero at $t = +600$ fs. In addition to revealing the rise time of the evanescent field to be much less than one picosecond, the sequence of images in Fig. 2b also shows that the field extends at the interface to ~50 nm (1/e) into vacuum on either side of the nanotube. The image length scale is consistent with theoretical considerations of optically excited plasmons[25]. The time necessary for the image counts to decay from the maximum to minimum values, normalized by the maximum change in counts per unit time (Fig. 2a, bottom) is 130 fs, reflecting the rate of change, as discussed below. A movie of the time dependent behaviour of the carbon nanotubes is included in the Supplementary Information.

The ultrafast response shown in the discrete energy gain and loss (that is, bands on both sides of the ZLP; Fig. 1) can be quantified in energy and time space (Fig. 3). In order to obtain the intensity profile, each energy spectrum was fitted to a series of Gaussians having the form:

$$S(E) = \frac{1-2\alpha}{\sqrt{2\pi\sigma^2}} e^{-E^2/2\sigma^2} + \sum_{\pm n} \frac{a_n}{\sqrt{2\pi\sigma^2}} e^{-(E\pm n\hbar\omega)^2/2\sigma^2} \quad (1)$$

where $\alpha$ is a sum over $a_n$, the amplitude of the $n$th photon process, and $\sigma$ reflects the energy width. A typical fit of equation (1), which has also been invoked in photoelectron studies[8], to the observed spectrum is shown in Fig. 1b. In Fig. 3a, the temporal dependence of different sidebands and the ZLP is plotted on a log scale. With Gaussian analysis in the time domain, we obtained the time constants involved. For the $\pm 3\hbar\omega$ peaks, $\tau_\sigma = 420$ fs, which is a direct result of the convolution of the femtosecond excitation pulse, the electron packet, and response time of the evanescent optical field; the femtosecond optical pulse duration $\tau_p = 220$ fs (ref. 11) and energy-filtering may be significant in reducing electron pulse energy dispersion[26]. All peaks were fitted similarly, giving the range of $\tau_\sigma$ to be $510 \pm 90$ fs. Because of the geometry involving the two pulses used in UEM-2, the (axial) group velocity mismatch is irrelevant here, as it results in a dispersion of ~1 fs over the 100 nm path length.

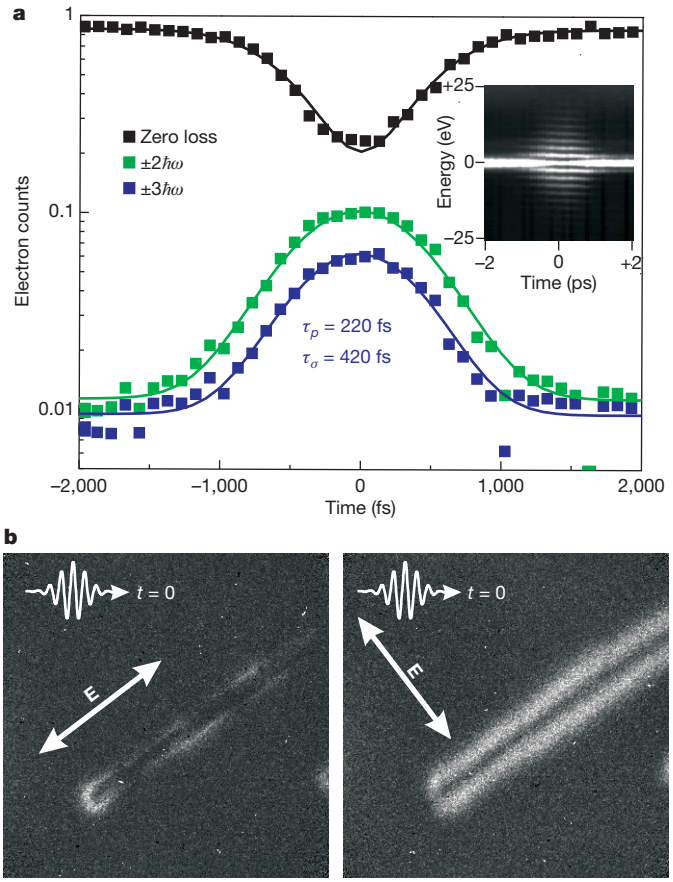



**Figure 3 | Temporal response and polarization dependence of the imaged interfacial fields. a**, Temporal dependence of the ZLP and representative peaks of $\pm n\hbar\omega$ plotted on a log scale. Given are the values for $\tau_p$ of the femtosecond optical pulse and $\tau_\sigma$ of the fitted transient; see text. The $t = 0$ position is determined by the temporal response of the ZLP[12]. Inset, linear contour plot of the data fitted by the Gaussians. **b**, Two images taken when the *E*-field polarization of the femtosecond laser pulse is parallel to (left image) and perpendicular to (right image) the long-axis of the nanotube. Both polarization frames were taken at $t = 0$, when the interaction between electron, photons and the evanescent field is at a maximum.

Besides the temporal and energy domain observations, we also examined the spatial distribution of the near field from images taken for different polarizations of the femtosecond pulse relative to the orientation of an individual nanotube (or wire). Shown in Fig. 3b are UEM images obtained at $t = 0$ with the *E*-field of the femtosecond laser pulse polarized either parallel (left frame) or perpendicular (right frame) to the long axis of the carbon nanotube. These frames display striking changes in the images: when the laser polarization is positioned appropriately relative to the nanotube orientation, a spatial enhancement of the evanescent field is observed in the UEM images. This is because for this case the confinement is in the regime of $d < \lambda$. On the other hand, for the other polarization, when $d > \lambda$, the tip enhancement is seen when the polarization changes by 90°. For this case, certain spatial modes may be formed with unique distributions (Fig. 3 and Supplementary Information). The precise distribution of the field is dependent upon the nanoscale geometry of the specimen[23,25], and the fact that the apex of the tip has a decrease in counts at perpendicular polarization is consistent with nanometric-scale field calculations[27].

The above experiments on carbon nanotubes were repeated using silver nanowires, obtaining similarly the filtered energy gain images, the electron spectra, and the polarization dependence (Supplementary Fig. 1). The irradiance needed, however, was an order of magnitude lower (10 GW cm$^{-2}$), consistent with the stronger near-field formed in the metallic nanowire and with the difference in material property.

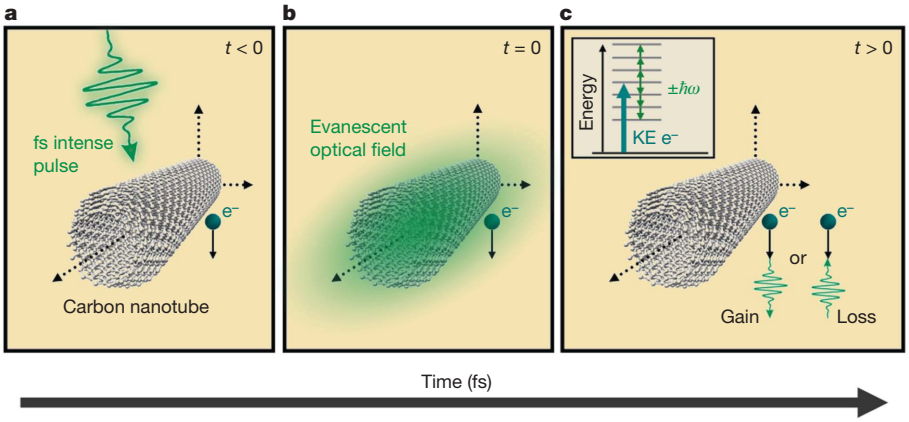



**Figure 4 | Physical depiction of the interaction between the electron, photon and the evanescent field. a**, A frame when the electron packet arrives at the nanotube before the femtosecond laser pulse ($t < 0$); no spatiotemporal overlap has yet occurred. **b**, The precise moment at $t = 0$ when the electron packet, femtosecond laser pulse and evanescent field are at maximum overlap at the carbon nanotube. **c**, Illustration of the process during and immediately after the interaction ($t > 0$) when the electron gains/loses energy equal to integer multiples of femtosecond laser photons. Inset, the possible final energies in the continuum due to the free–free transitions between the imaging electron and the photons in the femtosecond laser pulse. KE, kinetic energy.

The polarization dependence was the same as that of the carbon nanotubes, owing to the similarity of the geometrical structure of the nanotubes and nanowires.

From the above UEM results, the PIN picture can be illustrated by considering the spatiotemporal coordinates of the three bodies involved (Fig. 4). At negative times ($t < 0$), we can visualize the femtosecond laser pulse as not impinging on the nanostructure, and no discrete ($n\hbar\omega$) electron energy gain or loss would be observed. When the laser pulse encounters the nanostructure ($t = 0$), it creates the near-field excitations, and this interaction causes the surface field to oscillate with the electric field of the laser[22]. Because the nanotube diameter ($\sim$140 nm) is much less than the wavelength of the light (519 nm), the field is confined ($d < \lambda$) by the dimensions of the tube (wire), and this confinement sets up an oscillating dipole in the structure. The intensity of the evanescent field extends beyond the structure of the nanotube and falls off exponentially with distance from the surface[2,25]. Thus, evanescent fields effectively mediate the interaction between the 200 keV electron and the 2.4 eV photons in the femtosecond excitation pulse, but the absorption/emission processes only occur when both the electron and photon are overlapped in space at the nanostructure and in time at $t = 0$.

The orders of magnitude enhancement achieved in UEM may be appreciated when comparing with time-averaged, CW mode of excitation. For a tightly focused CW laser ($10^6$ W cm$^{-2}$), the number of excitations on the timescale of the field is nearly five orders of magnitude less than achieved in UEM using $\sim$100 GW cm$^{-2}$ irradiance. Further, for CW powers of about 10 W, it would be necessary for the nanostructures to dissipate the energy without significant structural damage. In UEM, typically the average power is of the order of 100 mW. Perhaps most importantly, the precise overlap of pulses in UEM allows for signal acquisition times of only a few seconds, as every electron contributes to the gain/loss signal on the timescale of the field's existence. In contrast, for CW electron spectroscopy, the signal will be overwhelmed by a background whose magnitude is conditioned by the repetition rate and other factors. Finally, we note that the process of $\pm n\hbar\omega$ absorption/emission reported here takes place for each single-electron, timed packet.

Photon-induced near-field imaging with electrons is made possible by the precise overlap of ultrafast electron packets, intense ultrafast laser pulses, and nanostructures. These structure-mediated (electron–photon interaction) phenomena, as well as the spatiotemporal properties of the evanescent electric fields, can now be imaged in real space and on the femtosecond timescale. By knowing the distribution of the field and the control over its polarization and temporal behaviour, it is possible to explore the nature of interfacial fields and their role in a variety of applications at the nanoscale of materials[28]. In optics, near-field imaging is now known to break the diffraction limit, but with limited spatial and temporal resolutions. In the present contribution, PINEM promises to take the resolutions into the domain of ultrafast electron microscopy, the atomic scale. Moreover, with PINEM, which exploits inelastic interactions, the real space images (and diffraction), which are the result of elastic interactions, can easily be obtained by removing the energy filter; the scanning requirement of optical near-field methods is not of concern here. Because the method involves surface electrons, it also can bring about imaging with sub-femtosecond electrons[13,29,30].

**Received 11 September; accepted 10 November 2009.**


1. Betzig, E. & Trautman, J. K. Near-field optics: microscopy, spectroscopy, and surface modification beyond the diffraction limit. *Science* **257,** 189–195 (1992).
2. Maier, S. A. & Atwater, H. A. Plasmonics: localization and guiding of electromagnetic energy in metal/dielectric structures. *J. Appl. Phys.* **98,** 011101 (2005).
3. Spence, J. C. H. *High-Resolution Electron Microscopy* (Oxford Univ. Press, 2003).
4. Kapitza, P. L. & Dirac, P. A. M. The reflection of electrons from standing light waves. *Proc. Camb. Phil. Soc.* **29,** 297–300 (1933).
5. Freimund, D. L., Aflatooni, K. & Batelaan, H. Observation of the Kapitza-Dirac effect. *Nature* **413,** 142–143 (2001).
6. Bucksbaum, P. H., Schumacher, D. W. & Bashkansky, M. High-intensity Kapitza-Dirac effect. *Phys. Rev. Lett.* **61,** 1182–1185 (1988).
7. Krausz, F. & Ivanov, M. Attosecond physics. *Rev. Mod. Phys.* **81,** 163–234 (2009).
8. Saathoff, G., Miaja-Avila, L., Aeschlimann, M., Murnane, M. M. & Kapteyn, H. C. Laser-assisted photoemission from surfaces. *Phys. Rev. A* **77,** 022903 (2008).
9. Howie, A. Electrons and photons: exploiting the connection. *Inst. Phys. Conf. Ser.* **161,** 311–314 (1999).
10. García de Abajo, F. J. & Kociak, M. Electron energy-gain spectroscopy. *N. J. Phys.* **10,** 073035 (2008).
11. Barwick, B., Park, H. S., Kwon, O.-H., Baskin, J. S. & Zewail, A. H. 4D imaging of transient structures and morphologies in ultrafast electron microscopy. *Science* **322,** 1227–1231 (2008).
12. Carbone, F., Kwon, O.-H. & Zewail, A. H. Dynamics of chemical bonding mapped by energy-resolved 4D electron microscopy. *Science* **325,** 181–184 (2009).
13. Zewail, A. H. & Thomas, J. M. *4D Electron Microscopy: Imaging in Space and Time* (Imperial College Press, 2009).
14. Midgley, P. A., Saunders, M., Vincent, R. & Steeds, J. W. Energy-filtered convergent-beam diffraction: examples and future prospects. *Ultramicroscopy* **59,** 1–13 (1995).
15. Boersch, H., Geiger, J. & Stickel, W. Interaction of 25-keV electrons with lattice vibrations in LiF. Experimental evidence for surface modes of lattice vibration. *Phys. Rev. Lett.* **17,** 379–381 (1966).
16. Korte, K. E., Skrabalak, S. E. & Xia, Y. Rapid synthesis of silver nanowires through a CuCl- or $CuCl_2$ polyol process. *J. Mater. Chem.* **18,** 437–441 (2008).
17. Park, H. S., Baskin, J. S., Kwon, O.-H. & Zewail, A. H. Atomic-scale imaging in real and energy space developed in ultrafast electron microscopy. *Nano Lett.* **7,** 2545–2551 (2007).
18. Dravid, V. P. *et al.* Buckytubes and derivatives: their growth and implications for buckyball formation. *Science* **259,** 1601–1604 (1993).

19. Ishikawa, R., Bae, J. & Mizuno, K. Energy modulation of nonrelativistic electrons in an optical near field on a metal microslit. *J. Appl. Phys.* **89,** 4065–4066 (2001).
20. Muller, H. G., van Linden van den Heuvell, H. B. & van der Wiel, M. J. Dressing of continuum states after MPI of Xe in a two-colour experiment. *J. Phys. At. Mol. Opt. Phys.* **19,** L733–L739 (1986).
21. Agostini, P., Fabre, F., Mainfray, G., Petite, G. & Rahman, N. K. Free-free transitions following six-photon ionization of xenon atoms. *Phys. Rev. Lett.* **42,** 1127–1130 (1979).
22. Kim, S. *et al.* High-harmonic generation by resonant plasmon field enhancement. *Nature* **453,** 757–760 (2008).
23. Burda, C., Chen, X., Narayanan, R. & El-Sayed, M. A. Chemistry and properties of nanocrystals of different shapes. *Chem. Rev.* **105,** 1025–1102 (2005).
24. Hamada, N., Sawada, S. & Oshiyama, A. New one-dimensional conductors: graphitic microtubules. *Phys. Rev. Lett.* **68,** 1579–1581 (1992).
25. Kawata, S., Inouye, Y. & Verma, P. Plasmonics for near-field nano-imaging and superlensing. *Nature Photon.* **3,** 388–394 (2009).
26. Baum, P. & Zewail, A. Femtosecond diffraction with chirped electron pulses. *Chem. Phys. Lett.* **462,** 14–17 (2008).
27. Novotny, L., Bian, R. X. & Xie, X. S. Theory of nanometric optical tweezers. *Phys. Rev. Lett.* **79,** 645–648 (1997).
28. Humphreys, C. J. *Understanding Materials* (Maney Publishing, 2002).
29. Baum, P. & Zewail, A. H. Attosecond electron pulses for 4D diffraction and microscopy. *Proc. Natl Acad. Sci. USA* **104,** 18409–18414 (2007).
30. Veisz, L. *et al.* Hybrid DC-AC electron gun for fs-electron pulse generation. *N. J. Phys.* **9,** 451 (2007).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** This work was supported by the National Science Foundation and the Air Force Office of Scientific Research in the Gordon and Betty Moore Center for Physical Biology at the California Institute of Technology. We thank S. Skrabalak for synthesizing and providing the silver nanowires.

**Author Contributions** All authors contributed extensively to the work presented in this paper.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.H.Z. (zewail@caltech.edu).

LETTERS

# Fault zone fabric and fault weakness

Cristiano Collettini[1], André Niemeijer[2]†, Cecilia Viti[3] & Chris Marone[2]

**Geological and geophysical evidence suggests that some crustal faults are weak[1–6] compared to laboratory measurements of frictional strength[7]. Explanations for fault weakness include the presence of weak minerals[4], high fluid pressures within the fault core[8,9] and dynamic processes such as normal stress reduction[10], acoustic fluidization[11] or extreme weakening at high slip velocity[12–14]. Dynamic weakening mechanisms can explain some observations; however, creep and aseismic slip are thought to occur on weak faults, and quasi-static weakening mechanisms are required to initiate frictional slip on mis-oriented faults, at high angles to the tectonic stress field. Moreover, the maintenance of high fluid pressures requires specialized conditions[15] and weak mineral phases are not present in sufficient abundance to satisfy weak fault models[16], so weak faults remain largely unexplained. Here we provide laboratory evidence for a brittle, frictional weakening mechanism based on common fault zone fabrics. We report on the frictional strength of intact fault rocks sheared in their *in situ* geometry. Samples with well-developed foliation are extremely weak compared to their powdered equivalents. Micro- and nanostructural studies show that frictional sliding occurs along very fine-grained foliations composed of phyllosilicates (talc and smectite). When the same rocks are powdered, frictional strength is high, consistent with cataclastic processes. Our data show that fault weakness can occur in cases where weak mineral phases constitute only a small percentage of the total fault rock and that low friction results from slip on a network of weak phyllosilicate-rich surfaces that define the rock fabric. The widespread documentation of foliated fault rocks along mature faults in different tectonic settings and from many different protoliths[4,17–19] suggests that this mechanism could be a viable explanation for fault weakening in the brittle crust.**

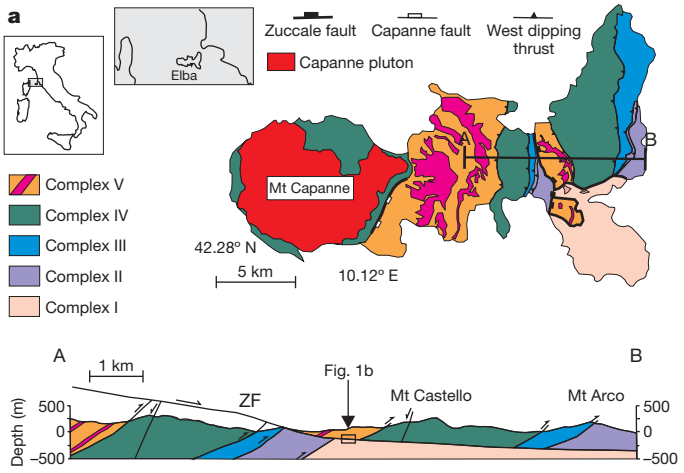
Laboratory measurements on a wide variety of rock types show that fault friction $\mu$ is in the range 0.6–0.8, independent of rock type, with the exception of a few weak minerals[7]. Several lines of evidence suggest that Byerlee's friction with $\mu = 0.6$ is applicable to many faults within the brittle crust[20], including *in situ* stress measurements showing that failure occurs on optimally oriented faults with normal friction and nearly hydrostatic pore pressure[21]. Moreover, the absence of moderate-to-large earthquakes along severely misoriented faults in both extensional and compressional environments is consistent with laboratory friction values[22]. In contrast, a number of studies have suggested that some mature crustal faults are weak compared to laboratory friction values[1–6,8,9]. Dynamic weakening mechanisms may explain some of these data. High speed friction experiments have shown that friction decreases above a threshold slip rate[12–14] and such weakening is consistent with observations of dynamic rupture[10,11,23]. However, a number of observations indicate that fault creep, aseismic slip and slip on unfavourably oriented faults with respect to the applied stress occur at low resolved shear stress, which implies that some faults are statically weak[1–4,6,8,9], that is, friction is low in the long term including both co-seismic and inter-seismic phases.

To investigate the role of fabrics in long term fault weakening, we conducted friction experiments on fault rocks obtained from the Zuccale fault, which is a low-angle normal fault exposed on the Isle of Elba (Fig. 1) in central Italy. Geological data suggest that the Zuccale fault has accommodated a total shear displacement of 6–8 km along a dip angle of about 15° in a stress field with a vertical maximum compression; that is, the fault was weak[24]. The fault zone is characterized by a foliated and lithologically heterogeneous fault core several metres thick. The adjacent hangingwall and footwall rocks deformed predominantly by brittle cataclastic processes. Within the fault core, however, a significant amount of deformation was accommodated within highly foliated phyllosilicate-rich horizons that represent deformation processes that occurred at less than 8 km of crustal depth[24]. The foliated microstructure probably formed in the early geological stages of fault activity by dissolution and precipitation processes. Once formed, the phyllosilicate-rich network deforms predominantly by frictional sliding along the phyllosilicate foliae[24].

We chose rock samples from two foliated zones (Fig. 1) for detailed friction studies (see Methods). Zone L2 is composed of massive calcite-rich portions and calcite sigmoids dispersed within a fine-grained ($<2\,\mu m$) foliation made of tremolite and phyllosilicate (smectite, talc and minor chlorite). Zone L3 contains both large crystals of calcite and agglomeration of tremolite fibres in random orientation, within a foliation made of fine-grained tremolite and phyllosilicates (smectite, talc and minor chlorite). We collected undisturbed samples of the fault rocks and cut them into wafers 0.8–1.2 cm thick and 5 cm × 5 cm in area. We refer to these as solid experiments. We also carried out experiments on powders obtained from crushing and sieving ($<150\,\mu m$) the solid samples. With this approach we directly compare the frictional properties of foliated fault rocks and their powdered equivalents, which have identical mineralogical composition (Table 1).

Frictional sliding experiments were carried out in the double direct shear configuration (Fig. 2a inset) at 25 °C and over a range of normal stresses from 10 to 150 MPa and shear slip velocities of 1 to 300 $\mu m\,s^{-1}$ (see Methods). All experiments are characterized by an initial strain hardening, where the shear stress increases rapidly during elastic loading, before a yield point, followed by shear at a steady-state value of frictional stress (Fig. 2a). We measured the steady state, residual, frictional shear stress for intact fault rocks and powders at each normal stress. Each rock type plots along a line consistent with a brittle failure envelope (Fig. 2b), but the solid wafers are much weaker than their powdered analogues. In particular, the powders show friction of about 0.6, whereas the foliated rocks have significantly lower values (0.45–0.20). We note that at each normal stress, foliated fault rocks have a friction coefficient that is 0.2–0.3 lower than the powders made from them (Fig. 3). The majority of our experiments were conducted

[1]Geologia Strutturale e Geofisica, Dipartimento di Scienze della Terra Università degli Studi di Perugia, 06100, Perugia, Italy. [2]Department of Geosciences and Energy Institute Center for Geomechanics, Geofluids, and Geohazards, Penn State University, University Park, Pennsylvania 16802, USA. [3]Dipartimento di Scienze della Terra Università degli Studi di Siena, 53100, Siena, Italy. †Present address: Istituto Nazionale di Geofisica e Vulcanologia, 00143, Roma, Italy.



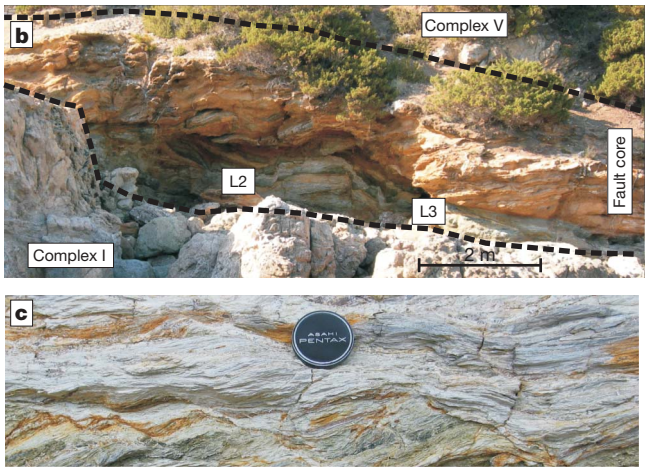




**Figure 1 | Example of a foliated low-angle normal fault.** **a**, Map and cross-section of the Zuccale fault (ZF) as exposed on the Isle of Elba, Central Italy. The Zuccale fault is an exhumed, ancient low-angle normal fault belonging to the Apenninic system. In the tectonically active area of the Apennines, to the east, a similar low-angle normal fault is producing microseismicity at depths[3] of 4–16 km. The geological cross-section AB through central and eastern Elba highlights the low-angle geometry[24]. **b**, Outcrop photograph of the fault structure and locations of zones L2 and L3 where samples were collected. **c**, Detail of the foliated fault rock in L2. (Camera lens cap shown for approximate scale; diameter 4 cm.)

without water and at room temperature, but we also ran experiments under water-saturated conditions and these indicate further weakening for intact fault rocks (for example, Fig. 3).

Microstructural studies of the tested rocks show that the sliding surfaces of the foliated solid wafers are located along the pre-existing very-fine-grained foliation made of tremolite and phyllosilicates (Fig. 4a and c). Transmission electron microscope (TEM) images show that the foliation is made of tremolite clasts surrounded by an interconnected network of phyllosilicates, including smectite and talc (Fig. 4e and f). The through-going microstructure is affected by translation and rotation of the (001) phyllosilicate layers with frequent interlayer delaminations, resulting in talc grain size reduction, down to 20 nm in thickness. This anastomosing network of phyllosilicate lamellae appears to provide a multitude of possible slipping planes for frictional shear under a low stress. In contrast, our experiments conducted on powder composed of the same materials indicate that much of the deformation occurs along zones characterized by grain-size reduction and affected by shear localization along R1, Y and B shears (Fig. 4b, for example[25,26]). The slipping zones of powdered samples are characterized by abundant calcite clasts immersed in a groundmass consisting of fine-grained tremolite and subordinate phyllosilicates (Fig. 4d).

Although the foliated wafers of intact fault rock and their powders have identical mineralogical compositions, the foliated samples are much weaker than their powdered analogues. On the basis of microstructural studies, we propose that weakness of the foliated fault rocks is due to the reactivation of pre-existing fine-grained and phyllosilicate-rich surfaces that are absent in the powders. The slightly higher friction coefficient of fault zone L2 compared to L3 (0.31 versus 0.25) is probably due to the presence of foliation-perpendicular calcite-filled fractures that reduce the interconnectivity of the microstructure. The frictional strength of the solid wafers is comparable to that of pure talc at similar sliding conditions[16] even with the presence of 65–80% of strong calcite and tremolite minerals up to

**Table 1 | Experimental details and mineral composition of the rocks tested.**

| Experiment | Sample | Normal stress (MPa) | Mineralogy (%) | | | | |
|---|---|---|---|---|---|---|---|
| | | | Calcite | Tremolite | Smectite | Talc | Chlorite |
| P2045 | Wafer L2 | 10; 20; 35; 50 | 43 | 36 | 14 | 6 | 1 |
| P2048 | Wafer L3 | 10; 20; 35; 50 | 39 | 26 | 19 | 15 | 1 |
| P2049 | Powder L3 | 10; 20; 35; 50 | 39 | 26 | 19 | 15 | 1 |
| P2050 | Powder L3 | 75;100;150 | 39 | 26 | 19 | 15 | 1 |
| P2052 | Wafer L3 | 75;100;150 | 39 | 26 | 19 | 15 | 1 |
| P2053 | Wafer L2 | 75;100;150 | 43 | 36 | 14 | 6 | 1 |
| P2054 | Wafer L2 | 10; 20; 35; 50 | 43 | 36 | 14 | 6 | 1 |
| P2056 | Powder L2 | 10; 20; 35; 50 | 43 | 36 | 14 | 6 | 1 |
| P2057 | Wafer-wet L2 | 10; 20; 35; 50 | 43 | 36 | 14 | 6 | 1 |
| P2066 | Powder L2 | 75;100;150 | 43 | 36 | 14 | 6 | 1 |

Mineral abundances were evaluated by TG-MS and XRPD (see Methods and Supplementary Fig. 1).

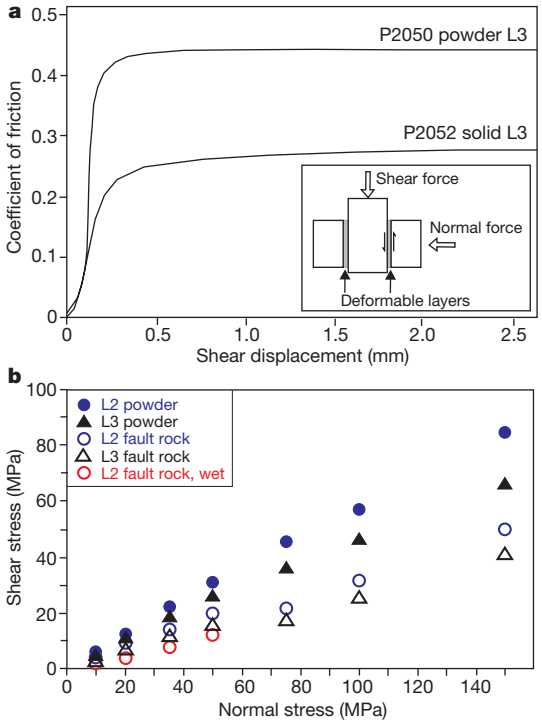



**Figure 2 | Friction experiments.** **a**, The inset shows the double-direct shear geometry used to shear solid and powdered samples of fault rock. The main plot shows the coefficient of friction $\mu$ versus shear displacement at the layer boundaries for a representative experiment at a normal stress of 150 MPa (sample L3). Strain hardening occurs initially, after which friction reaches a steady-state value. We note that sliding friction is significantly lower for the solid sample than for its powdered equivalent. **b**, Steady-state shear strength, measured during frictional sliding, plotted as a function of normal stress. Data for layers of powdered rock plot along lines with slopes $\mu = 0.55$ and $\mu = 0.43$ for L2 and L3 respectively. Data for shear of solid fault rock plot along lines with slopes $\mu = 0.31$ and $\mu = 0.25$ for L2 and L3 respectively.

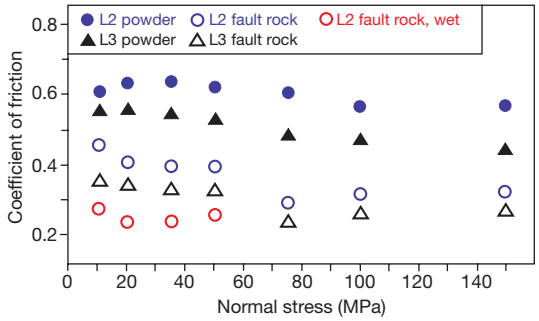



**Figure 3 | Frictional properties of fault rocks and powders made from them.** Steady-state friction plotted as a function of normal stress. The solid samples show friction values that are consistently 0.2–0.3 weaker than powders. Further weakening for intact fault rocks is achieved under water-saturated conditions.

200 μm wide (Table 1). This suggests that weak fault rocks can form with even small amounts of phyllosilicates as long as they form an interconnected, through-going foliated network. Although our experiments show that further weakening along phyllosilicate-rich rocks can be achieved by (1) the presence of fluids within the fault zone and (2) slow sliding velocities, the average value of $\mu = 0.25$ from our dry experiments at room temperature is sufficient to explain the absence of measurable heat flow along weak faults and frictional reactivation of faults oriented up to 75° from the maximum compressive stress, like the San Andreas or the low-angle normal faults of the Apennines.

In the core of ancient exhumed fault zones, where much of the deformation is accommodated, geological observations suggest that pervasive fluid influx can facilitate the dissolution of strong minerals and the precipitation of weak and fine-grained secondary phases[2,6,17,24,27]. With incremental shearing, these weak phyllosilicates can align to form a foliation that ultimately results in an interconnected network that allows slip to be accommodated along the weak foliation and reduces the frictional strength of the fault zones[2,6,17,24,27]. These observations have been confirmed in laboratory experiments using rock analogue materials where the combination of fluid-assisted deformation and the development of a phyllosilicate foliation at low sliding velocities and high shear strains caused major weakening of the experimental fault gouge (a decrease in $\mu$ from 0.8 to 0.25)[28,29]. Here, by performing frictional sliding experiments on intact wafers of natural fault rocks with a pre-existing phyllosilicate-rich foliation, we have demonstrated that the presence of a phyllosilicate foliation in fault zones results in significant fault weakening.

One important question involves whether fault weakening via fabric development in phyllosilicate rich rocks rules out earthquake-like frictional instability. Previous works have noted the tendency for weak materials to exhibit inherently stable frictional slip[15,16,26]. However, geological investigations have documented the mutual superposition between slip on phyllosilicates and brittle (hydrofractures) or earthquake-related structures (pseudotachylytes)[30]. Sharp and continuous slip zones of highly sheared clay-rich fault gouge have been interpreted as seismogenic principal displacement zones[13]. Continuous strands of phyllosilicates usually bound lenses of stronger lithologies[18,19]: these lenses could represent sites for stress concentrations and earthquake nucleation near patches of fault creep[18,23,31]. In addition, laboratory experiments on gouge mixtures show that when the phyllosilicates constitute a small percentage of the fault (10–30%), velocity-weakening behaviour is favoured at higher slip velocities[29,31]. A similar weak and velocity-weakening behaviour has been documented for smectite clay at low normal stress[32] (<30 MPa) and for clay-rich fault gouge in high-velocity friction experiments[14]. In this view, some crustal faults can behave as weak structures over long timescales (millions of years) and be intermittently seismogenic on shorter timescales.

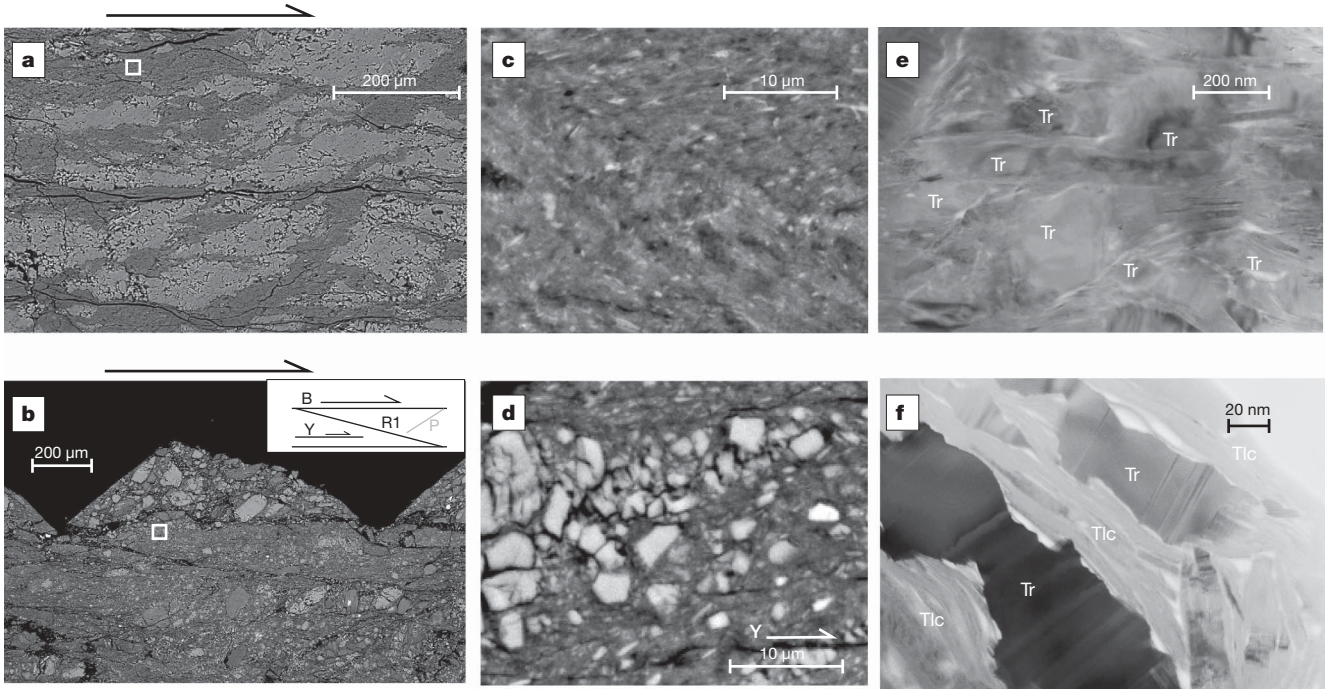



**Figure 4 | Comparison between solid-foliated and powder sliding surfaces in L3.** **a**, **c**, Solid experiment (P2048), total shear displacement of ~25 mm, steady friction coefficient $\mu = 0.32$. The calcite sigmoids are interdispersed within a fine-grained foliation (see detail in **c**) made of tremolite and phyllosilicates (smectite, chlorite and talc). The slipping processes occur along planes localized within the foliation. **b**, **d**, Powder experiment (P2049), total shear displacement ~25 mm, steady friction coefficient $\mu = 0.53$. The fault rock shows a cataclastic texture with zones affected by grain-size reduction (see detail in **d**) and shear localization along the R1, Y and B sliding surfaces. The little boxes in **a** and **b** represent the areas of **c** and **d**, respectively. The half arrows above **a** and **b** show direction of shear. **a**–**d** are scanning electron microscopy back-scattered electron images. **e**, **f**, TEM images, showing details of the foliated microstructure (solid experiment P2048) that is made of an interconnected network of thin, oriented phyllosilicates (smectite and talc; low-contrast portions) surrounding tremolite fibres. Tr, tremolite; Tlc, talc.

We show that fault weakening depends strongly on rock fabric and the distribution of weak phases within a fault zone. The result is akin to findings from studies involving dissolution[6] and ductile deformation[33]. However, we show that fabric-induced weakening can occur in the cataclastic, brittle deformation regime via frictional processes. Previous works focused on brittle fault strength have looked primarily at homogeneous mineral mixtures and bulk composition of the fault zone. Fluid–rock interactions and the growth of interconnected, phyllosilicate-rich networks have been documented in a wide variety of rock types[4,17–19,24,27,30]. We therefore propose that frictional sliding along such rock fabrics may be a viable mechanism to explain the mechanics of weak faults within the brittle crust.

## METHODS SUMMARY

To determine the mineralogical composition of the fault rocks (Table 1), we have quantitatively combined results from X-ray powder diffraction (XRPD) and thermo-gravimetry/mass spectrometry (TG-MS) studies.

For friction experiments we collected large blocks of foliated fault rocks pertaining to L2 and L3 domains and we cut them to form wafers 0.8–1.2 cm thick and 5 cm × 5 cm in area. The wafers were oriented so that they could be sheared in their *in situ* orientation, with foliation parallel to shear direction. We refer to these as solid experiments. We also carried out experiments on powders obtained from crushing and sieving the solid wafers. Experiments were conducted in a servo-controlled biaxial deformation apparatus[16,26] (Fig. 2a inset), at room temperature and over a range of shear slip velocities from 1 to 300 $\mu m\,s^{-1}$. Data reported on Figs 2 and 3 refer to a sliding velocity of 10 $\mu m\,s^{-1}$.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 4 May; accepted 12 October 2009.**


1. Zoback, M. D. *et al.* New evidence on the state of stress of the San Andreas fault system. *Science* **238,** 1105–1111 (1987).
2. Holdsworth, R. E. Weak faults—rotten cores. *Science* **303,** 181–182 (2004).
3. Chiaraluce, L., Chiarabba, C., Collettini, C., Piccinini, D. & Cocco, M. Architecture and mechanics of an active low-angle normal fault: Alto Tiberina Fault, northern Apennines, Italy. *J. Geophys. Res.* **112,** B10310, doi:10.1029/2007JB005015 (2007).
4. Moore, D. E. & Rymer, M. Talc-bearing serpentinites and the creeping section of the San Andreas fault. *Nature* **448,** 795–797, doi:10.1038/nature06064 (2007).
5. Brune, J. N., Henyey, T. L. & Roy, R. F. Heat flow, stress, and rate of slip along the San Andreas Fault, California. *J. Geophys. Res.* **74,** 3821–3827 (1969).
6. Wintsch, R. P., Christoffersen, R. & Kronenberg, A. K. Fluid-rock reaction weakening of fault zones. *J. Geophys. Res.* **100,** 13021–13032 (1995).
7. Byerlee, J. D. Friction of rocks. *Pure Appl. Geophys.* **116,** 615–629 (1978).
8. Rice, J. R. in *Fault Mechanics and Transport Properties of Rocks* (eds Evans, B. & Wong, T.-f.) 475–503 (Academic Press, 1992).
9. Faulkner, D. R., Mitchell, T. M., Healy, D. & Heap, M. J. Slip on 'weak' faults by the rotation of regional stress in the fracture damage zone. *Nature* **444,** 922–925 (2004).
10. Ampuero, J.-P. & Ben-Zion, Y. Cracks, pulses and macroscopic asymmetry of dynamic rupture on a bimaterial interface with velocity-weakening friction. *Geophys. J. Int.* **173,** 674–692 (2008).
11. Melosh, H. J. Dynamical weakening of faults by acoustic fluidization. *Nature* **279,** 601–606 (1996).
12. Di Toro, G., Hirose, T., Nielsen, S., Pennacchioni, G. & Shimamoto, T. Natural and experimental evidence of melt lubrication of faults during earthquakes. *Science* **311,** 647–649 (2006).
13. Wibberley, C. A. J. & Shimamoto, T. Earthquake slip weakening and asperities explained by thermal pressurization. *Nature* **436,** 689–692 (2005).
14. Boutareaud, S. *et al.* Clay-clast aggregates: A new textural evidence for seismic fault sliding? *Geophys. Res. Lett.* **35,** L05302, doi:10.1029/2007GL032554 (2008).
15. Scholz, C. H. *The Mechanics of Earthquakes and Faulting* 2nd edn, 1–508 (Cambridge University Press, 2002).
16. Carpenter, B. M., Marone, C. & Saffer, D. Frictional behavior of materials in the 3D SAFOD volume. *Geophys. Res. Lett.* **36,** L05302, doi:10.1029/2008GL036660 (2009).
17. Vrolijk, P. & van der Pluijm, B. A. Clay gouge. *J. Struct. Geol.* **21,** 1039–1048 (1999).
18. Faulkner, D. R., Lewis, A. C. & Rutter, E. H. On the internal structure and mechanics of large strike-slip faults: field observations from the Carboneras fault, southeastern Spain. *Tectonophysics* **367,** 235–251 (2003).
19. Jefferies, S. P. *et al.* The nature and importance of phyllonite development in crustal-scale fault cores: an example from the Median Tectonic Line, Japan. *J. Struct. Geol.* **28,** 220–235 (2006).
20. Scholz, C. H. Evidence for a strong San Andreas fault. *Geology* **28,** 163–166 (2000).
21. Townend, J. & Zoback, M. D. How faulting keeps the crust strong. *Geology* **28,** 399–402 (2000).
22. Collettini, C. & Sibson, R. H. Normal faults, normal friction? *Geology* **29,** 927–930 (2001).
23. Noda, H., Dunham, E. M. & Rice, J. R. Earthquake ruptures with thermal weakening and the operation of major faults at low overall stress levels. *J. Geophys. Res.* **114,** B07302, doi:10.1029/2008JB006143 (2009).
24. Collettini, C., Viti, C., Smith, S. A. F. & Holdsworth, R. E. The development of interconnected talc networks and weakening of continental low-angle normal faults. *Geology* **37,** 567–570 (2009).
25. Beeler, N. M., Tullis, T. E., Blanpied, M. L., &. Weeks, J. D. Frictional behavior of large displacement experimental faults. *J. Geophys. Res.* **101,** 8697–8715 (1996).
26. Marone, C. Laboratory-derived friction laws and their application to seismic faulting. *Annu. Rev. Earth Planet. Sci.* **26,** 643–696 (1998).
27. Evans, J. P. & Chester, F. M. Fluid rock interaction in faults of the San Andreas system: inferences from San Gabriel fault-rock geochemistry and microstructures. *J. Geophys. Res.* **100,** 13007–13020 (1995).
28. Bos, B., Peach, C. J. & Spiers, C. J. Frictional-viscous flow of simulated fault gouge caused by the combined effects of phyllosilicates and pressure solution. *Tectonophysics* **327,** 173–194 (2000).
29. Niemeijer, A. R. & Spiers, C. J. Velocity dependence of strength and healing behaviour in simulated phyllosilicate-bearing fault gouge. *Tectonophysics* **427,** 231–253 (2006).
30. Imber, J. *et al.* in *The Internal Structure of Fault Zones: Implications for Mechanical and Fluid-Flow Properties* (eds Wibberley, C. A. J. *et al.*) Vol. 299, 151–173 (Geological Society of London Special Publication, 2008).
31. Niemeijer, A. R. & Spiers, C. J. A microphysical model for strong velocity weakening in phyllosilicate-bearing fault gouges. *J. Geophys. Res.* **112,** B10405, doi:10.1029/2007JB005008 (2007).
32. Saffer, D. M., Frye, K. F., Marone, C. & Mair, K. Laboratory results indicating complex and potentially unstable frictional behavior of smectite clay. *Geophys. Res. Lett.* **28,** 2297–2300 (2001).
33. Shea, W. T. J. & Kronenberg, A. K. Strength and anisotropy of foliated rocks with varied mica contents. *J. Struct. Geol.* **15,** 1097–1121 (1993).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank I. Faoro for cutting the samples and J. P. Ampuero, D. Faulkner, R. Holdsworth and S. Smith for discussions. This research was motivated in part by stimulating discussions with P. Montone, M. Barchi and M. Cocco. We gratefully acknowledge funding by NSF grants OCE-0196462 EAR-0510182 and an INGV-DPC S5 M. Barchi grant. A.N. was supported in part by the ERC St. G. Nr.205175 USEMS project.

**Author Contributions** C.C., A.N. and C.M. designed the study. A.N. and C.C. carried out the experiments. A.N., C.C. and C.M. conducted the data analysis. C.C. and C.V. carried out the microstructural studies. C.V. did TEM and mineralogical characterization. All the authors contributed to the writing.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.C. (colle@unipg.it).

## METHODS

We chose rock samples from two highly foliated domains of the Zuccale fault (L2 and L3 in Fig. 1). Mineral abundances in selected fault rocks have been determined by coupling XRPD and TG-MS. For each sample, we collected three different XRPD patterns from:
(1) powders (grain size $<150\,\mu m$), representative of the bulk sample mineralogy (massive carbonate fragments + foliation);
(2) fine fraction (grain size $<2\,\mu m$, oriented samples), representative of the fine-grained foliation;
(3) fine fractions after glycolation, to distinguish expandable versus not-expandable clays.

The XRPD patterns on fine fractions revealed smectite, tremolite and talc as the main minerals within the foliation (Supplementary Fig. 1). The relative amounts of smectite, talc and tremolite, as reported in Table 1, were determined on the basis of XRPD intensity ratios. The total amount of calcite in the bulk samples was determined by TG-MS. The TG-MS data indicated that calcite decomposition occurred in the temperature range 600–800 °C. From the weight losses registered in this temperature range, we determined total calcite contents of 43% and 39% in L2 and L3, respectively.

For the experiments we collected blocks of cohesive foliated fault rocks pertaining to L2 and L3 domains and we cut them to form wafers 0.8–1.2 cm thick and 5 cm × 5 cm in area. The wafers were oriented so that they could be sheared in their *in situ* orientation, with foliation parallel to shear direction. We refer to these as solid experiments. We also carried out experiments on powders obtained from crushing and sieving ($<150\,\mu m$): (1) intact pieces of fault rock and (2) the samples used in the solid experiments. With this approach we directly compare the frictional properties of foliated fault rocks and their powdered equivalents, which have identical mineralogical compositions (Table 1). We found no difference between experiments conducted on layers of powdered fault rock and powders from solid experiments.

Experiments were conducted in a double direct shear geometry using a servo-controlled biaxial deformation apparatus[16,26] (Fig. 2a inset). Three steel blocks sandwiched two layers of fault rock while the horizontal hydraulic ram, in load feedback mode, applied normal load to the layers. The vertical ram was advanced in displacement feedback mode at constant velocity, thereby shearing the two fault rock layers. Both normal and shear loads were measured by load cells at the load point with a precision of 0.1 kN. Displacement of the horizontal and vertical axes was measured using displacement transducers with 0.1 μm precision. The sheared layers were initially 0.8–1.2 cm thick (solid fault rock) or 0.6 cm thick (powders) and we measured changes in layer thickness continuously during shear. Frictional contact area is constant during shear. Loads and displacements were continuously recorded for each axis at 10 kHz and averaged to rates from 1 to 100 Hz, depending on the sliding velocity. Normal and shear stresses were obtained by dividing the measured applied loads by the contact area of the forcing block (5 × 5 cm) in the case of normal stress, and twice the contact area in the case of shear stress.

LETTERS

# Common ecology quantifies human insurgency

Juan Camilo Bohorquez[1], Sean Gourley[2], Alexander R. Dixon[3], Michael Spagat[4] & Neil F. Johnson[2]

**Many collective human activities, including violence, have been shown to exhibit universal patterns[1–19]. The size distributions of casualties both in whole wars from 1816 to 1980 and terrorist attacks have separately been shown to follow approximate power-law distributions[6,7,9,10]. However, the possibility of universal patterns ranging across wars in the size distribution or timing of within-conflict events has barely been explored. Here we show that the sizes and timing of violent events within different insurgent conflicts exhibit remarkable similarities. We propose a unified model of human insurgency that reproduces these commonalities, and explains conflict-specific variations quantitatively in terms of underlying rules of engagement. Our model treats each insurgent population as an ecology of dynamically evolving, self-organized groups following common decision-making processes. Our model is consistent with several recent hypotheses about modern insurgency[18–20], is robust to many generalizations[21], and establishes a quantitative connection between human insurgency, global terrorism[10] and ecology[13–17,22,23]. Its similarity to financial market models[24–26] provides a surprising link between violent and non-violent forms of human behaviour.**

The political scientist Spirling[27] and others[9,10] have correctly warned that finding common statistical distributions (for example, power laws) in sociological data is not the same as understanding their origin. Possible political, ideological, cultural, historical and geographical influences make conflict arguably one the 'messiest' of all human activities to analyse. Mindful of these challenges, yet inspired by recent studies of human dynamics[1–11,17,28], we analyse the size and timing of 54,679 violent events reported within nine diverse insurgent conflicts, placing equal emphasis on both finding and modelling common patterns. Such insurgencies typify the future wars and threats faced by society[18,19].

Our data sources are real-time media databases, official (government and non-governmental organization) reports and academic studies. Supplementary Information provides details, plus data-set extracts. The event data from different conflicts were compiled by different researchers, often with cross-checking by independent research teams, thereby reducing systematic collection or recording biases. Comparison of event accounts across a wide range of sources[12,29] reduces potential media bias and mistaken aggregation (for example, misreporting two events of sizes $x_1$ and $x_2$ as one event of size $x_3 = x_1 + x_2$), which would create significant errors in a tail-dependent estimate such as a power-law slope. We focus on measuring deaths, because injuries are harder to cross-check. However, where possible, we check that our conclusions are robust to the inclusion of injuries.

Figures 1 and 2 show our empirical findings for event size, whereas Fig. 3 shows event timings. Our model (described later and shown schematically in Fig. 4) provides a quantitative explanation of these findings by treating the insurgent population as an ecology of dynamically evolving, decision-making groups, in line with several recent sociological hypotheses[18–20]. In addition to explaining the ubiquity of approximate power laws in the event size distribution and the apparent central role of the 2.5 exponent value (Fig. 1), it explains the conflict-dependent deviations beyond a power law (see green curves in Fig. 2). Furthermore, the same model framework also explains the common burstiness in the distribution of event timings that we observe across insurgent conflicts (see black curves in Fig. 3).

Figure 1 gives exponent values, obtained by applying Clauset *et al.*'s[9,10] established methodology for estimating discrete power-law distributions, $p(x) \approx x^{-\alpha}$, for $x \geq x_{\min}$ where $x_{\min}$ is estimated together with $\alpha$. In all cases we cannot reject the hypothesis that the size distribution of the events follows a power law, but we can reject log-normality. Four detailed examples are shown in Fig. 2. Following our preliminary 2005 results for Iraq and Colombia, we had suggested[12] that other insurgent wars might be clustered around $\alpha = 2.5$. All the insurgent wars that we have analysed support this hypothesis. By contrast, we find that the Spanish Civil War and the American Civil War—neither of which are considered insurgent—each give distributions where log-normal can not be rejected, and where even the best-fit $\alpha$ value is much smaller (near 1.7, which is the value for the aggregated sizes of conventional wars[9]). This finding provides quantitative support for claims circulating in social science[18,19]

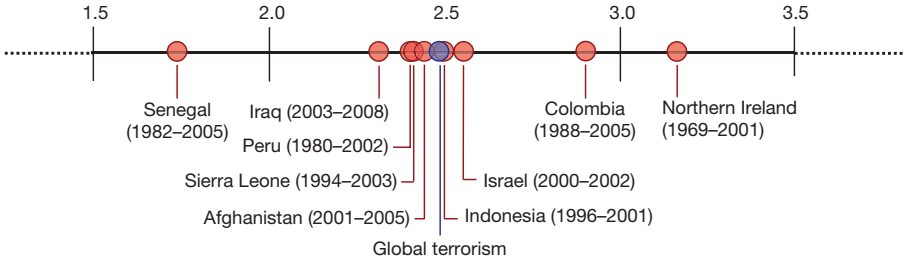



**Figure 1 | Power-law exponents.** Value $\alpha$ for power law $p(x) \approx x^{-\alpha}$ deduced from the empirical distributions of event size $x$ (that is, the number of casualties) for insurgent conflicts. Statistical procedures follow refs 9 and 10. Blue dot shows the value 2.48 for distribution of total size of global terrorist events, from Clauset *et al.*[10]. The years in parentheses describe the empirical data set range used to deduce $\alpha$, not the actual conflict duration.

[1]Department of Industrial Engineering and CEIBA Complex Systems Research Center, Universidad de Los Andes, Bogota, Colombia. [2]Complex Systems Group, Physics Department, University of Miami, Florida 33126, USA. [3]Cavendish Laboratory, Cambridge University, Cambridge CB3 0HE, UK. [4]Department of Economics, Royal Holloway College, University of London, Egham, Surrey TW20 0EX, UK.

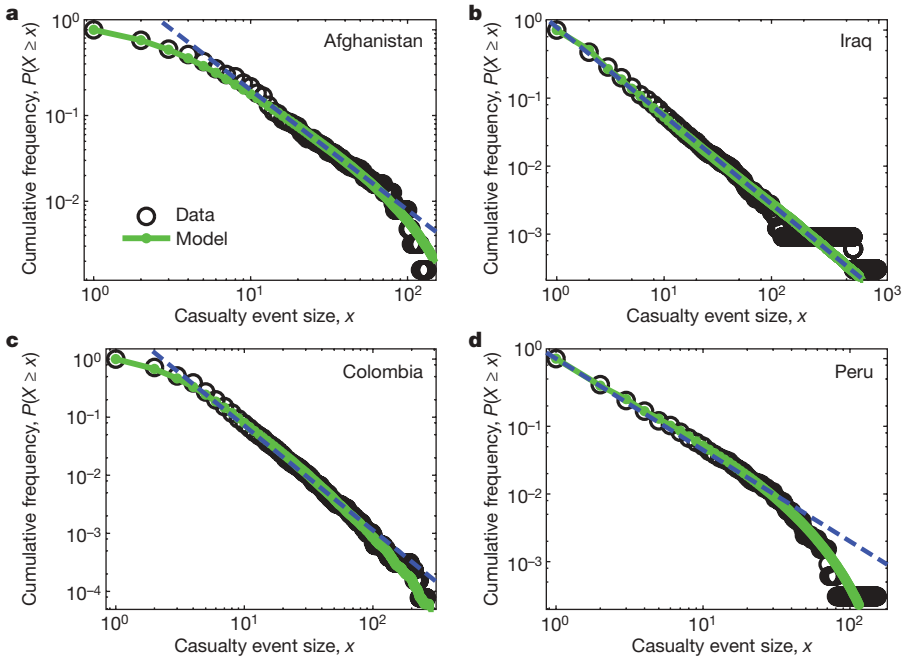



**Figure 2 | Size of events. a–d**, Log–log plot of the complementary cumulative distribution of event size $P(X \geq x)$ (that is, the probability of an event of size greater than or equal to $x$) for four conflicts from Fig. 1. Horizontal axis shows event size $x$, namely the number of casualties. Solid green curves show the results from our model. Blue dashed line is a straight line guide to the eye, not a power-law fit.

that insurgent wars represent 'open-source'[18], 'fourth-generation'[19] warfare, with qualitatively different dynamics from traditional wars. Several trivial explanations of the data can be ruled out, such as proportionality to city size[10].

Figure 3 demonstrates a common burstiness in the distribution for the number of events per day, $n$, irrespective of size. As explained in the Methods and Supplementary Information, we compare the distributions over daily event counts for different epochs within the four modern conflicts for which we have such data, against control distributions ('random war') obtained by randomizing event occurrences within each epoch. The data for each conflict (green circles) deviate from its random war (dashed curve) in a similar way: the real war exhibits an overabundance of light days (that is, days with few attacks) and of heavy days (that is, days with many attacks), but a 'lack' of medium days compared with the random war (see lower panel). By considering subsets of days, we have determined that these features are not just an artefact of a variation in attack volume across days of the week (for example, Fridays; see Supplementary Information). Interestingly, this

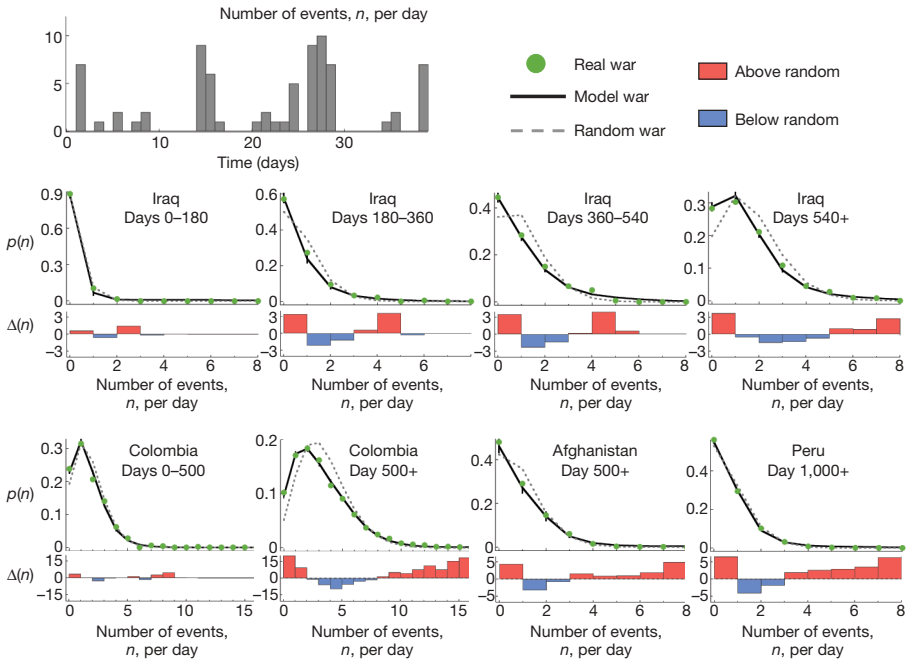



**Figure 3 | Timing of events.** A time series with $n = 0, 1, 2, \ldots$ events per day. Green circles show distribution $p(n)$ for the number of days with $n$ events in actual conflict. Dashed lines represent average values for random wars. Solid lines denote average distributions calculated from 10,000 realizations of our model (Fig. 4). Histograms below represent differences $\Delta(n)$ between real and random wars, in units of standard deviations from the mean. Error bars for random wars, namely one standard deviation from the mean of 10,000 shufflings, are shown but are small. Error bars for model wars demonstrate a small spread in run outcomes.

burstiness has become more pronounced over time for the wars in both Iraq and Colombia, suggesting that they have become less random as they have evolved. These findings are insensitive to the precise specification of the epochs within a given conflict.

Our model framework in Fig. 4 incorporates two key features: (1) ongoing group dynamics within the insurgent population (for example, as a result of internal interactions and/or the presence of an opposing entity such as a state army); (2) group decision-making about when to attack based on competition for media attention. Within this framework, we find that mechanism (1) dictates the features observed in Figs 1 and 2 whereas mechanism (2) dictates those in Fig. 3. Mechanism (1) is consistent with recent work on human group dynamics in everyday environments[28], and with current views of modern insurgencies as fragmented, transient and evolving[18]. Mechanism (2) is consistent with comments by former US Senior Counterinsurgency Adviser David Kilcullen, who noted[20] that when insurgents ambush an American convoy in Iraq, '... they're not doing that because they want to reduce the number of Humvees we have in Iraq by one. They're doing it because they want spectacular media footage of a burning Humvee.' We consider the insurgent population as having an overall strength $N$ comprising human combatants, information, resources and weapons—though, for simplicity, one can think of $N$ humans. $N$ is continually being repartitioned through coalescence and fragmentation processes, thereby producing an ecology of groups. A group's strength at time-step $t$ determines the number of human casualties $x$ it would produce if it decided to engage in an event at that time-step. We take $N$ to be approximately constant over time, though our main conclusions are unchanged if $N$ evolves slowly with small fluctuations.

These two coexisting dynamic mechanisms generate rich time series that can explain the numbers of events of different sizes at each time-step. However, because the data in Figs 1 and 2 are time-aggregated whereas those in Fig. 3 are size-aggregated, we can provide far more insightful explanations by using simplified versions that treat the respective non-dominant mechanism in an averaged way. Consider first the simple situation in which the group coalescence and fragmentation processes in the insurgent population are represented by probabilities[18]. The fragmentation probability $v_{\mathrm{frag}}$ is taken to be small ($\sim$1%) to mimic the infrequent situation in which a group member suddenly senses imminent danger and the entire group scatters. If fragmentation does not occur, the group may coalesce with another group with probability $v_{\mathrm{coal}}$. This mimics the situation in which two individuals initiate a communications link between them of arbitrary range (for example, a mobile phone call), and hence their respective groups of strength $s_1$ and $s_2$ act in a coordinated way with strength $(s_1 + s_2)$. Because these two processes can be triggered by any particular constituent group member at any time, the probability that it affects a specific group should be proportional to $s$ (refs 24–26). Treating mechanism (2) in an averaged way, we assume that all groups are equally likely to be involved in an event over time. This is consistent with the time-averaged behaviour of the full decision-making model (see later). The time-averaged distribution of group strengths $s$ therefore acts like the distribution of event sizes $x$ (ref. 12), resulting in a steady-state approximate power-law distribution whose analytic solution $\alpha = 2.5$ (refs 25, 26) is within the empirical bounds of Clauset *et al.*'s total value of $2.48 \pm 0.07$ for global terrorism[10]. This analytically obtained theoretical value[25,26] of 2.5 is robust to many model generalizations[21,25,26] (for example, coalescence of multiple groups, fragmentation into groups larger than one), thereby offering an explanation for the observed bunching of the empirical values around 2.5 in Fig. 1.

Invoking a more realistic mechanism for grouping dynamics than simple probabilities (see Supplementary Information), we find that our model framework can explain not only the approximate power-law behaviour and central role of the 2.5 exponent (Fig. 1) but also the behaviours beyond power law observed in Fig. 2. Accounting explicitly for an opposing population (for example, state army) with total strength $N_{\mathrm{B}}$, the coalescence and fragmentation are now caused by the interactions between groups. The casualties produced by clashes between opposing groups can then be used to obtain the event size distributions (green curves in Fig. 2). Full details are given in the Methods and Supplementary Information, with the four model parameters for each conflict (total insurgent strength $N_{\mathrm{A}}$, total state strength $N_{\mathrm{B}}$ and casualty scales $C_{\mathrm{S}}$ and $C_{\mathrm{L}}$). When two opposing groups meet they fight, with some members of both groups killed and the smaller group fragmenting. $C_{\mathrm{S}}$ ($C_{\mathrm{L}}$) sets the scale for the

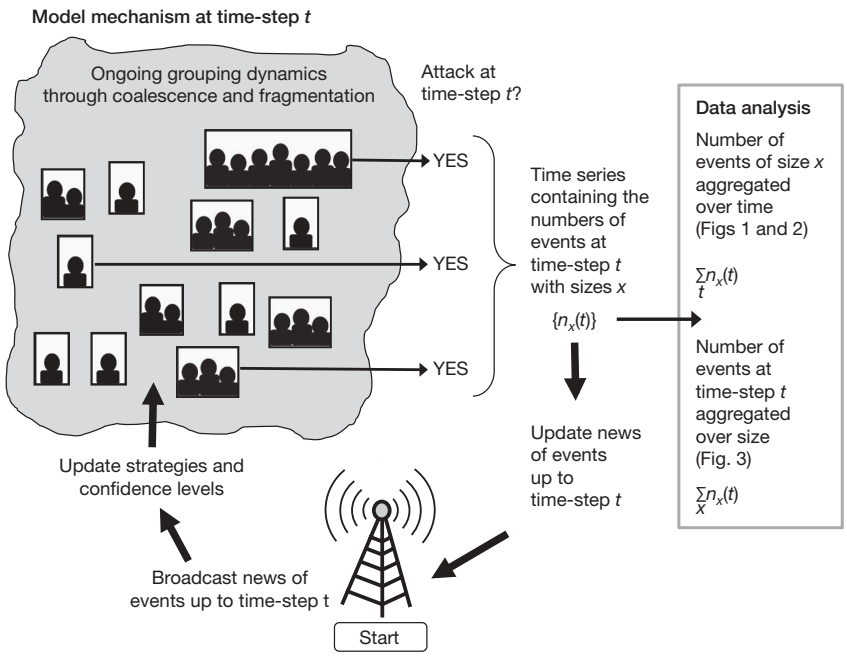



**Figure 4 | Model framework for insurgency.** The insurgent population comprises an overall strength $N$, distributed into groups with diverse strengths at each time-step $t$. This distribution changes over time as groups join and break up. Dark shadows indicate strength, and hence casualties that can be inflicted in an event involving that group. Figures 1 and 2 are derived from the number of events of size $x$ aggregated over time. Figure 3 is derived from the number of events at a given time-step aggregated over size.

number of the smaller (larger) group's members destroyed. As $C_S$ is increased, the model deviates increasingly from a straight line at low $x$, suggesting that Afghanistan and Colombia share the following similarity: in a clash in which the insurgent group is the smaller group, this insurgent group takes heavier relative losses than for the wars in Iraq and Peru. The ratio between the two populations' strengths ($N_A$ and $N_B$) tends to control the slope itself, with greater strength differences resulting in steeper slopes. This suggests that there might be a greater difference between the strengths of the army and insurgency in Colombia than in Iraq or Afghanistan. The total insurgent strength $N_A$ controls the large $x$ roll-off in Fig. 2. Afghanistan and Peru deviate substantially from power laws for large $x$, which our model interprets as relatively small insurgency strength. Colombia and Iraq hardly deviate from power laws for large $x$, implying greater insurgency strength.

Because Fig. 3 features data aggregated over size, we replace the detailed grouping dynamics (that is, mechanism (1)) by a time-averaged number of groups. Given the resolution of our data and the typical numbers of observed daily attacks, we take one time-step as equivalent to one day. If a group launches an attack during a day with many other attacks, its media coverage will in general be reduced. If, instead, it launches an attack on a quiet day, its media coverage will increase[20]. Each group receives daily some common but limited information (for example, public radio or newspaper announcements about previous attacks, opposition troop movements, a specific religious holiday, even a shift in weather patterns). The actual content is unimportant provided it becomes the primary input for the group's decision-making process. (See ref. 26 for a full description of an equivalent financial-market version.) Although the groups are heterogeneous in terms of their strategies, they tend to converge towards similar responses when fed the same information[26], thereby generating distributions (black curves) that are almost identical to those observed (green circles). Our model (see Supplementary Information and ref. 26) includes a confidence threshold that must be surpassed before any decision can be made, allowing us to interpret the increase in non-randomness over time for Iraq and Colombia as a decrease in this confidence threshold; that is, the insurgent groups in both wars have become less cautious over time about whether to launch attacks. Reference 30 presents independent empirical evidence that groups of humans do indeed use such generic decision-based mechanisms.

To our knowledge, our model provides the first unified explanation of high-frequency, intra-conflict data across human insurgencies. Other explanations of human insurgency are possible, though any competing theory would also need to replicate the results of Figs 1–3. Our model's specific mechanisms challenge traditional ideas of insurgency based on rigid hierarchies and networks, whereas its striking similarity to multi-agent financial market models[24–26] hints at a possible link between collective human dynamics in violent and non-violent settings[1–19].

## METHODS SUMMARY

For the event size distribution (Figs 1 and 2), we use Clauset *et al.*'s methodology[9,10] to estimate power-law exponents, and test power-law and log-normal hypotheses, with the time-aggregated time series of events. This methodology[9,10] is a widely accepted, published state-of-the-art statistical procedure for analysing power-law-like distributions. For the event timing distribution (Fig. 3), we divide the time series for the number of events per day into epochs. These epochs are chosen such that there is no significant trend in the moving-average within each epoch. The precise specification of each epoch's time-window does not affect our main findings. We then generate 10,000 random wars by shuffling the date of the events within each section, averaging across the shuffles. Our model (Fig. 4) replicates the empirical size and timing patterns of Figs 1–3. Full details are given in Methods and Supplementary Information.

**Received 5 July; accepted 29 October 2009.**


1. Gabaix, X., Parameswaran, G., Plerou, V. & Stanley, H. E. A theory of power law distributions in financial market fluctuations. *Nature* **423,** 267–270 (2003).
2. Lux, T. & Marchesi, M. Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* **397,** 498–500 (1999).
3. Dussutour, A., Fourcassie, V., Helbing, D. & Deneubourg, J. L. Optimal traffic organization in ants under crowded conditions. *Nature* **428,** 70–73 (2004).
4. Mantegna, R. N. & Stanley, H. E. Scaling behaviour in the dynamics of an economic index. *Nature* **376,** 46–49 (1995).
5. Barabasi, A. L. The origin of bursts and heavy tails in human dynamics. *Nature* **435,** 207–211 (2005).
6. Richardson, L. F. Variation of the frequency of fatal quarrels with magnitude. *J. Am. Stat. Assoc.* **43,** 523–546 (1948).
7. Cederman, L.-E. Modeling the size of wars: from billiard balls to sandpiles. *Am. Polit. Sci. Rev.* **97,** 135–150 (2003).
8. Lim, M., Metzler, R. & Bar-Yam, Y. Global pattern formation and ethnic/cultural violence. *Science* **317,** 1540–1544 (2007).
9. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51,** 661–703 (2009).
10. Clauset, A., Young, M. & Gleditsch, K. S. On the frequency of severe terrorist events. *J. Conflict Resolut.* **51,** 58–87 (2007).
11. Galam, S. & Mauger, A. On reducing terrorism power: a hint from physics. *Physica A* **323,** 695–704 (2003).
12. Johnson, N. *et al.* Universal patterns underlying ongoing wars and terrorism. Preprint at ⟨http://arxiv.org/abs/physics/0605035⟩ (2006).
13. Johnson, D. D. P. & Madin, J. S. in *Natural Security: A Darwinian Approach to a Dangerous World* (eds Sagarin, R. & Taylor, T.) Ch. 11 159–185 (Univ. California Press, 2009).
14. Flack, J. Security in an uncertain world. *Nature* **453,** 451–452 (2008).
15. Adams, E. S. & Mesterton-Gibbons, M. Lanchester's attrition models and fights among social animals. *Behav. Ecol.* **14,** 719–723 (2003).
16. Wrangham, R. W. in *Emerging Synthesis in Science* (ed. Pines, D.) 123–132 (Santa Fe Institute, 1985).
17. Epstein, J. M. *Nonlinear Dynamics, Mathematical Biology, and Social Science* (Addison-Wesley, 1997).
18. Robb, J. *Brave New War: The Next Stage of Terrorism and the End of Globalization* (Wiley, 2007).
19. Hammes, T. *The Sling and the Stone: On War in the 21st Century* (Zenith Press, 2004).
20. Packer, G. Knowing the enemy – can social scientists redefine the war on terror? *New Yorker* 18 December (2006).
21. Ruszczycki, B., Burnett, B., Zhao, Z. & Johnson, N. F. Relating the microscopic rules in coalescence-fragmentation models to the macroscopic cluster-size distribution. *Eur. Phys. J. B* doi:10.1140/epjb/e2009-00354-5 (in the press).
22. May, R. M. & McLean, A. R. *Theoretical Ecology Principles and Applications* 3rd edn (Oxford Univ. Press, 2007).
23. Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. Effective leadership and decision making in animal groups on the move. *Nature* **433,** 513–516 (2005).
24. Eguiluz, V. M. & Zimmermann, M. G. Transmission of information and herd behaviour: an application to financial markets. *Phys. Rev. Lett.* **85,** 5659–5662 (2000).
25. D'Hulst, R. & Rodgers, G. J. Exact solution of a model for crowding and information transmission in financial markets. *Int. J. Theor. Appl. Finance* **3,** 609–616 (2000).
26. Johnson, N. F., Jefferies, P. & Hui, P. M. *Financial Market Complexity* (Oxford Univ. Press, 2003).
27. Spirling, A. The next big thing: scale invariance in political science. ⟨http://www.people.fas.harvard.edu/~spirling/documents/powerlawSend.pdf⟩.
28. Palla, G., Barabási, A. L. & Vicsek, T. Quantifying social group evolution. *Nature* **446,** 664–667 (2007).
29. Hicks, M. H.-R. *et al.* The weapons that kill civilians — deaths of children and noncombatants in Iraq, 2003–2008. *N. Engl. J. Med.* **360,** 1585–1588 (2009).
30. Wang, W., Chen, Y. & Huang, J. Heterogeneous preferences, decision-making capacity, and phase transitions in a complex adaptive system. *Proc. Natl Acad. Sci. USA* **106,** 8423–8428 (2009).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We are grateful to J. Restrepo for his involvement in the data collection, and R. Zarama for discussions.

**Author Contributions** J.C.B., S.G., M.S. and N.F.J. worked on the data and data analysis. S.G., A.R.D. and N.F.J. worked on the model development. All authors participated in the writing and associated discussions, giving detailed feedback in all areas of the project.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.F.J. (njohnson@physics.miami.edu).

LETTERS

# Rapid formation and selective stabilization of synapses for enduring motor memories

Tonghui Xu[1]*, Xinzhu Yu[1]*, Andrew J. Perlik[1], Willie F. Tobin[1], Jonathan A. Zweig[1], Kelly Tennant[2], Theresa Jones[2] & Yi Zuo[1]

**Novel motor skills are learned through repetitive practice and, once acquired, persist long after training stops[1,2]. Earlier studies have shown that such learning induces an increase in the efficacy of synapses in the primary motor cortex, the persistence of which is associated with retention of the task[3–5]. However, how motor learning affects neuronal circuitry at the level of individual synapses and how long-lasting memory is structurally encoded in the intact brain remain unknown. Here we show that synaptic connections in the living mouse brain rapidly respond to motor-skill learning and permanently rewire. Training in a forelimb reaching task leads to rapid (within an hour) formation of post-synaptic dendritic spines on the output pyramidal neurons in the contralateral motor cortex. Although selective elimination of spines that existed before training gradually returns the overall spine density back to the original level, the new spines induced during learning are preferentially stabilized during subsequent training and endure long after training stops. Furthermore, we show that different motor skills are encoded by different sets of synapses. Practice of novel, but not previously learned, tasks further promotes dendritic spine formation in adulthood. Our findings reveal that rapid, but long-lasting, synaptic reorganization is closely associated with motor learning. The data also suggest that stabilized neuronal connections are the foundation of durable motor memory.**

Fine motor movements require accurate muscle synergies that rely on coordinated recruitment of intracortical synapses onto corticospinal neurons[6,7]. Obtaining new motor skills has been shown to strengthen the horizontal cortical connections in the primary motor cortex[4,5]. In this study, we taught mice a single-seed reaching task (Supplementary Movie 1). The majority of 1-month-old mice that underwent training gradually increased their reaching success rates during the initial 4 days, and then levelled off ($n = 42$, Fig. 1a, b). There were a few mice ($n = 5$) that engaged in extensive reaching, but continually failed to grasp the seeds. These mice normally gave up reaching after 4–8 days (Fig. 1b). To investigate the process of learning-induced synaptic remodelling in the intact motor cortex, we repeatedly imaged the same apical dendrites of layer V pyramidal neurons marked by the transgenic expression of yellow fluorescent protein (YFP-H line) in various cortical regions during and after motor learning, using transcranial two-photon microscopy[8] (Supplementary Fig. 1). Dendritic spines are the postsynaptic sites of most excitatory synapses in the brain and changes in spine morphology and dynamism serve as good indicators of synaptic plasticity[9,10]. Spines that were formed and eliminated were identified by comparing images from two time points, and then normalized to the initial images. Imaged regions were guided by stereotaxic measurements, ensuring the imaged neurons resided in the primary motor cortex. In several experiments, intracortical microstimulation was performed at the end of repetitive imaging to confirm that images were taken from the functionally responding motor cortex (Fig. 1c, Supplementary Notes and Supplementary Fig. 2).

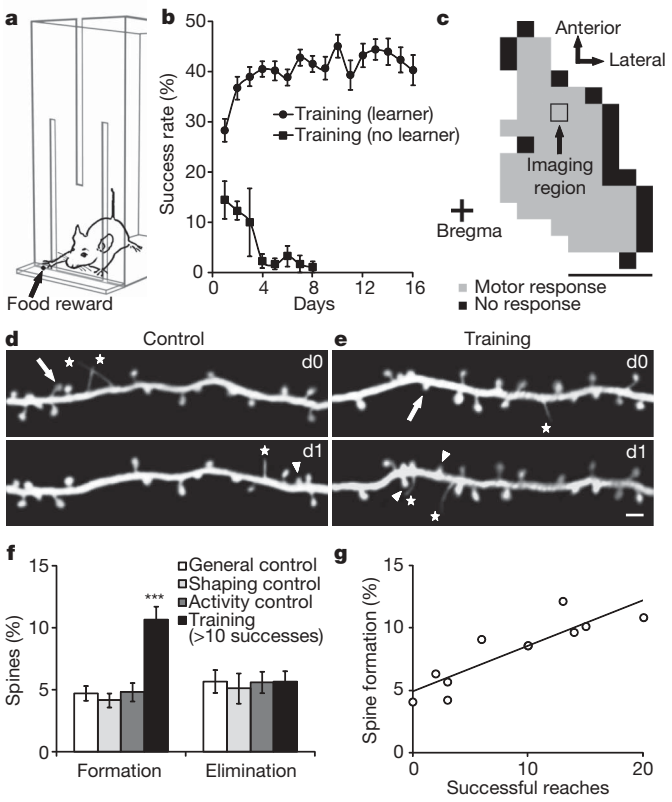



**Figure 1 | Motor skill learning in adolescent mice promotes immediate spine formation in the contralateral motor cortex. a**, A cartoon of motor training. **b**, Average success rates during training for learning and non-learning mice (mean ± s.e.m., 42 learners and 5 no learners). **c**, An intracortical microstimulation map indicates that the imaged region is within the motor cortex. Scale bar, 1 mm. **d**, **e**, Repeated imaging of the same dendritic branches over one-day intervals reveals spine elimination (arrows) and formation (arrowheads), and filopodia (asterisks) in a general control (**d**) and a trained (**e**) mouse. Scale bar, 2 μm. **f**, Percentage of spines formed and eliminated under various control and training conditions immediately following the first training session (mean ± s.d., $***P < 0.001$). **g**, The degree of spine formation observed following the first training session is linearly correlated with the number of successful reaches during this session ($r^2 = 0.77$).

[1]Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz, California 95064, USA. [2]Institute for Neuroscience, Department of Psychology, University of Texas at Austin, Austin, Texas 78712, USA.
*These authors contributed equally to this work.

Unexpectedly, we found that motor learning led to rapid formation of dendritic spines (spinogenesis) in the motor cortex contralateral to the reaching forelimb. One-month-old mice that finished 30 reaches with more than 10 successes in the first day of training were imaged within 1 h of the training session and showed 10.6 ± 1.1% new spines which were not in the images acquired the day before training. This spine formation was more than double that found in age-matched controls, which were handled similarly and imaged over the same period of time, but not trained (Fig. 1d–f, 4.7 ± 0.6% in general controls, $P < 0.001$). In contrast, spine elimination measured in the same images was not significantly altered by motor learning during single training sessions (Fig. 1f, $P > 0.9$). In addition, mice that went through shaping but not training (shaping controls) or mice that were trained to reach for a seed too far away to grasp (activity controls) did not show an increase in spine formation rates (Fig. 1f, $P > 0.1$ with general control, $P < 0.001$ with trained mice; see Methods for all control conditions). This suggests that refinement of fine motor movements, rather than other training-related experiences or unskilled motor activity, drives robust spine formation. Furthermore, the percentage of spines formed immediately after the first training session is linearly correlated with the number of successful reaches during the training session, revealing a direct link between learning and spine formation (Fig. 1g, $r^2 = 0.77$).

Perfection of a motor skill often requires persistent practice over time. To examine how prolonged learning affects spine dynamics, we trained and imaged mice over different periods of time (that is, from 2 to 16 days). We found that training for 2 days and longer resulted in significant increases, not only in spine formation, but also in spine elimination (Fig. 2a, b, $P < 0.005$ at all time points). Although delayed, this increase in spine elimination ultimately resulted in the total spine density in the trained animals returning to control levels by day 16 (Fig. 2c). As a control, we measured spine formation and elimination over a 4-day training period in the ipsilateral (to the trained limb) primary motor cortex and the contralateral posterior sensory cortex, and found no significant increase in spine formation or elimination in either case (Fig. 2a, b, d, e, $P > 0.2$). In addition, mice that failed to learn also failed to show an increase in either spine formation or elimination in the contralateral motor cortex (Figs 1b and 2a, b, f, $P > 0.6$). Therefore, the observed changes in spine dynamics are region- and learning-specific, indicating that motor learning causes synaptic reorganization in the corresponding motor cortex.

The enhanced spine loss after rapid spinogenesis reflects a rewiring of the neuronal circuitry in response to learning, rather than a simple addition of new spines. To examine how learning reorganizes synaptic connections, we imaged the same mice three times, classified imaged spines into new and pre-existing spines based on their appearance in the initial two images, and then quantified their survival percentages in the third images (Fig. 3a). Our data show that new spines are less stable than pre-existing spines in general (Fig. 3b, c). Specifically, in control mice, 43.8 ± 3.1%, 25.8 ± 5.3% and 19.2 ± 4.6% of the spines that formed between days 0 and 4 remained by days 6, 8 and 16, respectively. During the same period of time, 96.7 ± 0.5%, 94.9 ± 1.1% and 92.8 ± 1.9% of the pre-existing spines remained (Fig. 3d, $P < 0.001$ compared to new spines). These results suggest that new spines are initially unstable and undergo a prolonged selection process before being converted into stable synapses. In addition, we found that new spines were significantly more stable in trained mice, with 64.1 ± 2.2%, 55.3 ± 4.1% and 51.0 ± 4.8% of the spines that formed during the initial 4-day training remaining by days 6, 8 and 16, respectively (Fig. 3d, $P < 0.001$ compared to new spines in control mice). In contrast, pre-existing spines in trained mice were significantly less stable than control mice over the same time periods (Fig. 3d, $P < 0.05$). More importantly, when the fate of the new

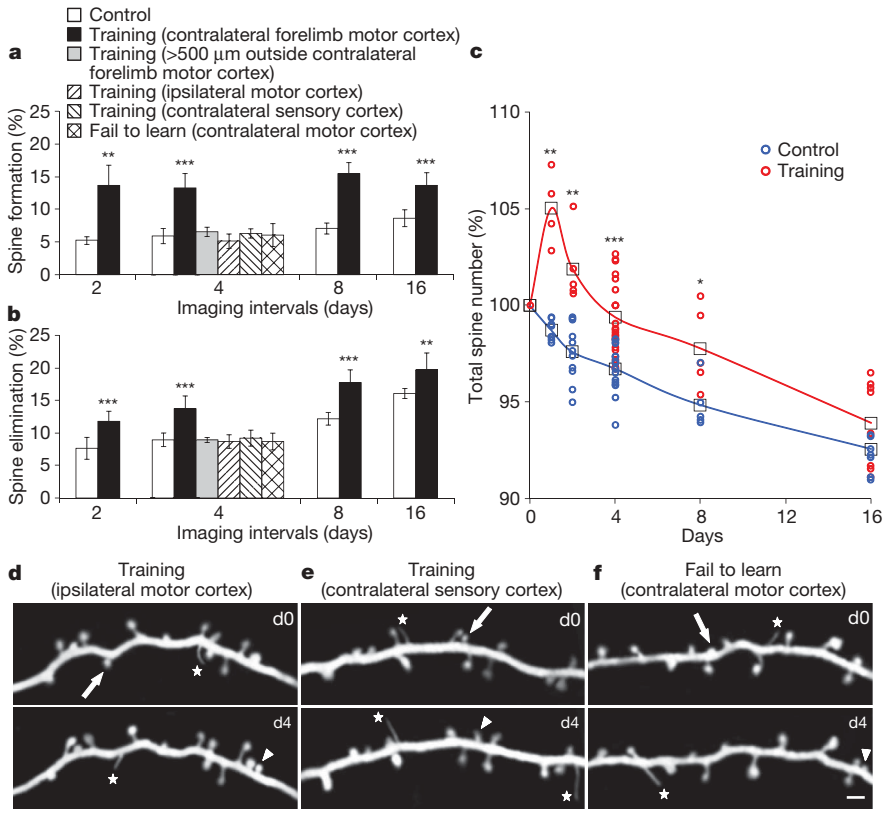



**Figure 2 | Enhanced spine dynamics during adolescent motor training is region- and learning-specific.** **a**, **b**, Percentage of spines formed (**a**) and eliminated (**b**) under control and training conditions. **c**, Total spine number increases during initial learning, but returns to normal levels with prolonged training. **d**, **e**, Imaging of the same dendritic branches over 4 days in the ipsilateral primary motor cortex (**d**) and the contralateral sensory cortex (**e**) of the trained mice. **f**, Imaging of the same dendritic branches over 4 days in the contralateral motor cortex of a mouse that failed to learn the task. Data are presented as mean ± s.d., $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. Scale bar, 2 μm.

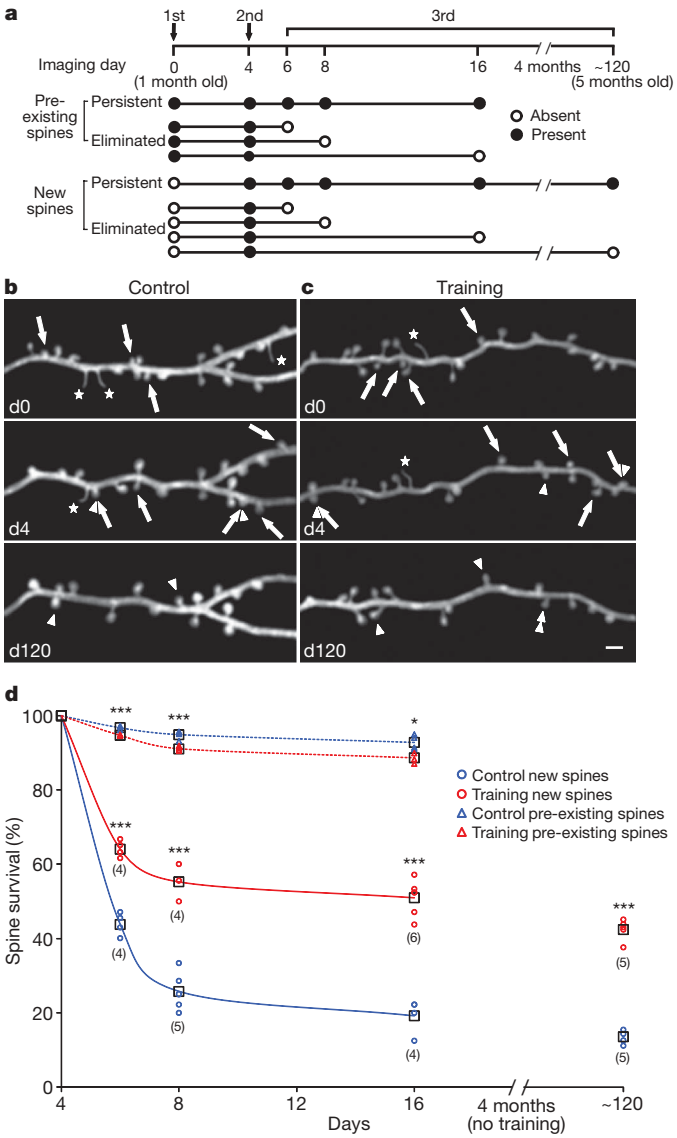



**Figure 3 | Motor skill learning stabilizes newly formed spines.** **a**, Timeline of experiments, showing possible outcomes. **b**, **c**, Repeated imaging of dendritic branches at 0, 4 and 120 days in a control (**b**) and a trained (**c**) mouse. Scale bar, 2 μm. **d**, Percentages of surviving new and pre-existing spines, as a function of time, for control and trained animals (mean ± s.d., $*P < 0.05$ and $***P < 0.001$). Numbers of animals examined at each time point are indicated below new spine data points.

spines formed during initial learning (day 0–4) was examined months later (day 120), we found that 42.3 ± 2.9% of new spines persisted in the mice trained for 16 days during adolescence, whereas only 13.5 ± 1.7% of new spines remained in the control mice (Fig. 3d, $P < 0.001$). In addition, we found that spine formation and stabilization were associated with behavioural improvement. More new spines were formed daily during the learning acquisition phase (days 1–4) than during the learning maintenance phase (days 5–16); the new spines that were formed during learning acquisition, but not during maintenance, were preferably stabilized with continuous training (Supplementary Notes and Supplementary Fig. 3). Taken together, these data indicate that motor learning selectively stabilizes learning-induced new spines and destabilizes pre-existing spines. The prolonged persistence of learning-induced synapses provides a potential cellular mechanism for the consolidation of lasting, presumably permanent, motor memories.

Dendrites in the mammalian brain contain not only spines but also filopodia. Filopodia are long, thin protrusions without bulbous heads, and make up ~10% of the total dendritic protrusions in the motor cortex of 1-month-old mice. Previous studies suggest that filopodia are precursors of dendritic spines[11,12]. We found that filopodia were very dynamic in the mouse motor cortex *in vivo*. Most of them turned over within 1 day in control mice (79.3 ± 12.8% formation and 87.6 ± 5.9% elimination), and motor learning had no effect on filopodial formation and elimination (91.0 ± 15.3% formation and 86.5 ± 8.8% elimination, $P > 0.2$). Among the filopodia observed in the initial images, few of them became spines over the following day in control mice (6.3%). However, this filopodium-to-spine transition was enhanced by motor skill learning (13.1%). Furthermore, 25% of new spines formed from filopodia on training day 1 persisted after another 4 days of training, indicating a contribution of filopodia to the rewired neuronal circuitry. Furthermore, when filopodia and spines were pooled together for analysis, there was a ~10% increase in the dynamics of both control and training categories. Thus, the conclusion of motor learning on total protrusions was consistent with the spine analysis alone (Supplementary Fig. 4).

One of the important characteristics of motor learning is that, once the skill is well learned, its further maintenance does not require constant practice. To test whether lasting motor memories might be contained within structurally stable neural circuits, we trained young mice for 8–16 days to acquire the reaching skill, housed them in control cage conditions for 4 months, and retrained them on the same task in adulthood. We found that these pre-trained mice maintained skilful performance with high success rates even on the first day of reintroducing the reaching task (Fig. 4a). Imaging of these pre-trained adult mice showed that spine formation and elimination during retraining were similar to those of naive adults without training (Fig. 4b, e, g, $P > 0.1$ for 4 and 8 days). In contrast, naive adults learning the reaching task for the first time had a learning curve similar to adolescent mice, and showed significantly higher spine formation and elimination than control adults (Fig. 4a–c, g, 4 and 8 days, $P < 0.01$ with control for both formation and elimination). Next, we asked if learning a novel motor skill continued to drive synaptic reorganization in the pre-trained brain. To do this, we trained mice that had been pre-trained on the reaching task with a new motor task—the capellini handling task, which also requires fine forelimb motor skills (see Methods). We found that pre-trained mice, similar to naive adults, had enhanced spine formation and elimination during the training of this novel skill task (Fig. 4d, f, g, $P < 0.001$ compared to control adults). Despite high spine dynamics induced by novel skill learning, most spines that were formed during adolescent learning of the reaching task and maintained in adults persisted after training with the capellini handling task (95.6 ± 7.7%), suggesting that already stabilized synapses are not perturbed by novel learning in adults. These results indicate that synaptic structural coding outlasts the early learning experience and persists in adulthood to support later maintenance of motor skills. The fact that novel learning experiences continue to drive synaptic reorganization without affecting the stability of synapses formed during previous learning further suggests that different motor behaviours are stored using different sets of synapses in the brain.

Our study investigated the process of synapse reorganization in the living brain during natural learning, distinguishing it from several studies where changes were triggered by non-physiological sensory manipulation[13–18]. Although rapid synapse formation has been observed during long-term potentiation *in vitro*[19,20], we show, for the first time, that synapse formation in the neocortex begins immediately as animals learn a new task in the living brain (within 1 h of training initiation). Such high spine formation does not occur with motor activity alone or later practice of the established skill. The rapidity of the response contradicts the general assumption that significant synaptic structural remodelling in motor cortex takes days to occur, following more subtle cellular activity and changes in synaptic efficacy[4,21,22]. One recent study on brain slices shows that glutamate-sensitive currents expressed in newly formed spines are indistinguishable from mature spines of comparable volumes[23], further suggesting

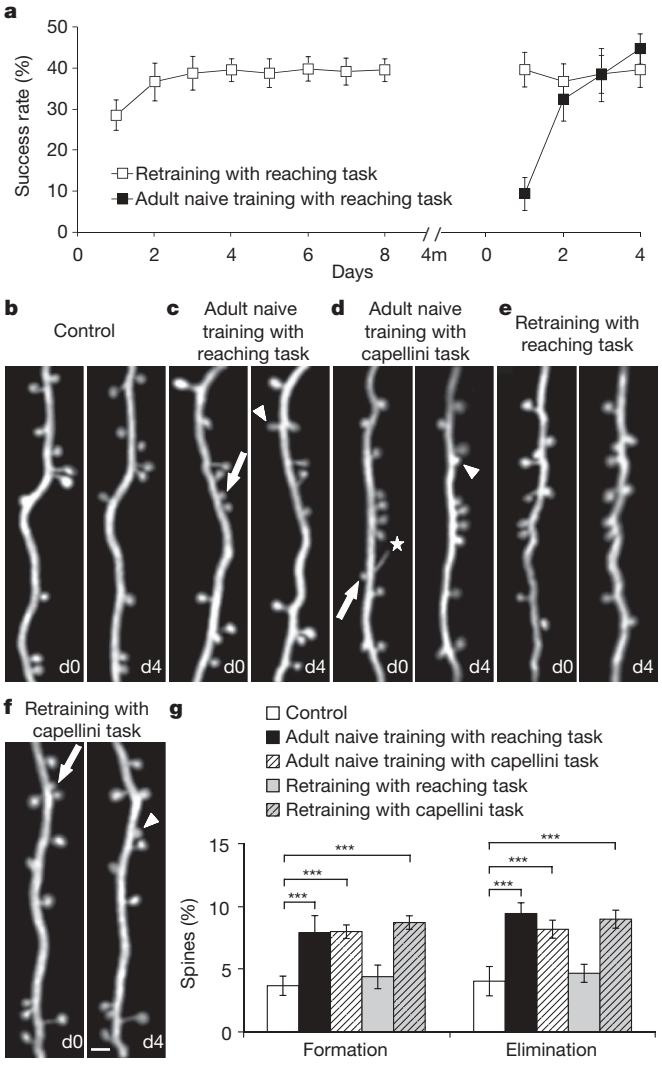



**Figure 4 | Novel motor skill training promotes spine formation and elimination in adult mice. a**, Pre-trained mice start with high success rates during adult retraining (mean ± s.e.m., 10 naive trained and 14 retrained adults). **b–f**, Repetitive imaging of dendritic branches over 4 days in a control adult (**b**), naive adults training with the reaching task (**c**) and capellini handling task (**d**), and pre-trained adults retraining with the same reaching task and the new capellini handling task (**f**). Scale bar, 2 μm. **g**, Percentages of spines formed and eliminated over 4 days in adult mice under different conditions (mean ± s.d., $***P < 0.001$).

that the new spines formed during learning are probably active. Furthermore, the persistence of new spines over months provides a long-lasting structural basis for the enhanced synaptic strength that is retained even when the task performance is discontinued.

Many previous studies have used fixed tissue preparation to investigate changes in synapse number and dendritic complexity after motor skill learning[24–28]. Our *in vivo* imaging of superficial dendrites from layer V pyramidal neurons revealed that postsynaptic dendritic spine addition was rapid, but eventually counteracted by the loss of pre-existing spines, resulting in a time-dependent spine density change during motor learning. Although the synaptogenesis observed in our study is compatible with earlier results, its temporal relationship with behavioural improvement and the contribution of synapse elimination in circuitry reorganization in other brain layers and regions during motor learning require further investigation. This eventual balancing of synapse number could be a homeostatic mechanism by which the output layer V neurons integrate converging inputs into superficial cortical layers to govern precise fine motor control.

## METHODS SUMMARY

Young (1 month old) and adult (>4 months old) mice expressing YFP in a small subset of cortical neurons (YFP-H line[29]) were used in all the experiments. Young mice were trained on the single-seed reaching task for up to 16 days and displayed a stereotypical learning curve (Fig. 1b). Naive adult mice and mice that had been previously trained with the single-seed reaching task in adolescence were trained with either the same reaching task or a novel capellini handling task for up to 8 days (see Methods). Apical dendrites of layer V pyramidal neurons, 10–100 μm below the cortical surface, were repeatedly imaged in mice under ketamine–xylazine anaesthesia with two-photon laser scanning microscopy. Spine dynamics in the motor cortex and other regions were followed over various intervals. Imaged regions were initially guided by stereotaxic measurements. In 14 mice, intracortical microstimulation (see Methods) was performed at the end of repetitive imaging to determine the location of acquired images relative to the functional forelimb motor map (Supplementary Fig. 2). In total, 32,079 spines from 209 mice were tracked over 2–4 imaging sessions, with 121 mice imaged twice, 79 mice three times and 9 mice imaged four times. Spine formation and elimination rates in each mouse were determined by comparing images of the same dendrites acquired at two time points; all changes were expressed relative to the total number of spines seen in the initial images. The number of spines analysed and the percentage of spine elimination and formation under various experimental conditions are summarized in Supplementary Table 1. To quantify spine size, calibrated spine head diameters were measured over time[30] (Supplementary Notes). All data are presented as mean ± s.d., unless otherwise stated. *P*-values were calculated using the Student's *t*-test. A non-parametric Mann–Whitney *U*-test was used to confirm all conclusions.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 7 April; accepted 6 August 2009.**
**Published online 29 November 2009.**

1. Luft, A. R. & Buitrago, M. M. Stages of motor skill learning. *Mol. Neurobiol.* **32,** 205–216 (2005).
2. Karni, A. *et al.* Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature* **377,** 155–158 (1995).
3. Rioult-Pedotti, M. S., Donoghue, J. P. & Dunaevsky, A. Plasticity of the synaptic modification range. *J. Neurophysiol.* **98,** 3688–3695 (2007).
4. Rioult-Pedotti, M. S., Friedman, D. & Donoghue, J. P. Learning-induced LTP in neocortex. *Science* **290,** 533–536 (2000).
5. Harms, K. J., Rioult-Pedotti, M. S., Carter, D. R. & Dunaevsky, A. Transient spine expansion and learning-induced plasticity in layer 1 primary motor cortex. *J. Neurosci.* **28,** 5686–5690 (2008).
6. Monfils, M. H., Plautz, E. J. & Kleim, J. A. In search of the motor engram: motor map plasticity as a mechanism for encoding motor experience. *Neuroscientist* **11,** 471–483 (2005).
7. Sanes, J. N. & Donoghue, J. P. Plasticity and primary motor cortex. *Annu. Rev. Neurosci.* **23,** 393–415 (2000).
8. Grutzendler, J., Kasthuri, N. & Gan, W. B. Long-term dendritic spine stability in the adult cortex. *Nature* **420,** 812–816 (2002).
9. Yuste, R. & Bonhoeffer, T. Morphological changes in dendritic spines associated with long-term synaptic plasticity. *Annu. Rev. Neurosci.* **24,** 1071–1089 (2001).
10. Gray, E. G. Electron microscopy of synaptic contacts on dendrite spines of the cerebral cortex. *Nature* **183,** 1592–1593 (1959).
11. Ziv, N. E. & Smith, S. J. Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* **17,** 91–102 (1996).
12. Dailey, M. E. & Smith, S. J. The dynamics of dendritic structure in developing hippocampal slices. *J. Neurosci.* **16,** 2983–2994 (1996).
13. Zuo, Y., Yang, G., Kwon, E. & Gan, W. B. Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* **436,** 261–265 (2005).
14. Trachtenberg, J. T. *et al.* Long-term *in vivo* imaging of experience-dependent synaptic plasticity in adult cortex. *Nature* **420,** 788–794 (2002).
15. Holtmaat, A., Wilbrecht, L., Knott, G. W., Welker, E. & Svoboda, K. Experience-dependent and cell-type-specific spine growth in the neocortex. *Nature* **441,** 979–983 (2006).
16. Hofer, S. B., Mrsic-Flogel, T. D., Bonhoeffer, T. & Hubener, M. Experience leaves a lasting structural trace in cortical circuits. *Nature* **457,** 313–317 (2009).
17. Keck, T. *et al.* Massive restructuring of neuronal circuits during functional reorganization of adult visual cortex. *Nature Neurosci.* **11,** 1162–1167 (2008).
18. Lendvai, B., Stern, E. A., Chen, B. & Svoboda, K. Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex *in vivo*. *Nature* **404,** 876–881 (2000).
19. Engert, F. & Bonhoeffer, T. Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399,** 66–70 (1999).
20. Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R. & Muller, D. LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402,** 421–425 (1999).
21. Kleim, J. A. *et al.* Cortical synaptogenesis and motor map reorganization occur during late, but not early, phase of motor skill learning. *J. Neurosci.* **24,** 628–633 (2004).

22. Adkins, D. L., Boychuk, J., Remple, M. S. & Kleim, J. A. Motor training induces experience-specific patterns of plasticity across motor cortex and spinal cord. *J. Appl. Physiol.* **101,** 1776–1782 (2006).
23. Zito, K., Scheuss, V., Knott, G., Hill, T. & Svoboda, K. Rapid functional maturation of nascent dendritic spines. *Neuron* **61,** 247–258 (2009).
24. Kleim, J. A., Vij, K., Ballard, D. H. & Greenough, W. T. Learning-dependent synaptic modifications in the cerebellar cortex of the adult rat persist for at least four weeks. *J. Neurosci.* **17,** 717–721 (1997).
25. Greenough, W. T., Larson, J. R. & Withers, G. S. Effects of unilateral and bilateral training in a reaching task on dendritic branching of neurons in the rat motor-sensory forelimb cortex. *Behav. Neural Biol.* **44,** 301–314 (1985).
26. Withers, G. S. & Greenough, W. T. Reach training selectively alters dendritic branching in subpopulations of layer II–III pyramids in rat motor-somatosensory forelimb cortex. *Neuropsychologia* **27,** 61–69 (1989).
27. Kleim, J. A. *et al.* Motor learning-dependent synaptogenesis is localized to functionally reorganized motor cortex. *Neurobiol. Learn. Mem.* **77,** 63–77 (2002).
28. Kolb, B., Cioe, J. & Comeau, W. Contrasting effects of motor and visual spatial learning tasks on dendritic arborization and spine density in rats. *Neurobiol. Learn. Mem.* **90,** 295–300 (2008).
29. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28,** 41–51 (2000).
30. Zuo, Y., Lin, A., Chang, P. & Gan, W. B. Development of long-term dendritic spine stability in diverse regions of cerebral cortex. *Neuron* **46,** 181–189 (2005).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank D. States, W. Thompson, L. Hinck, D. Feldheim, J. Ding, X. Li, A. Lin and C. Cirelli for critical comments on this manuscript; A. Sitko for her pilot studies of skilled reaching in mice, and D. Adkins, J. Kleim and N. Thomas for their assistance with intracortical microstimulation procedures. This work was supported by grants from the Ellison Medical Foundation, the DANA Foundation, and the National Institute on Aging to Y.Z.

**Author Contributions** T.X. and X.Y. contributed equally to this work. Both of them performed *in vivo* imaging, analysed the data, made figures and participated in the discussion. A.J.P., W.F.T. and J.A.Z. trained all the mice used in the experiments. K.T. and T.J. developed behavioural methods, performed the intracortical microstimulation experiments, and provided comments for the manuscript. Y.Z. initiated the project, did data analysis and wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.Z. (zuo@biology.ucsc.edu).

## METHODS

**Single-seed reaching task.** Mice were food-restricted to maintain 90% of free feeding weight before the start of training. The training chamber was constructed as a clear Plexiglas box 20 cm tall, 15 cm deep and 8.5 cm wide into which each individual mouse was placed. Three vertical slits 0.5 cm wide and 13 cm high were located on the front wall of the box: in the centre, on the left side, and on the right side (Supplementary Fig. 5). A 1.25-cm-tall exterior shelf was affixed to the wall in front of the slits to hold millet seeds for food reward. The training included two phases: 'shaping' and 'training'. The shaping phase (2–5 days in duration) was used to familiarize mice with the training chamber and task requirements and also to determine their preferred limbs. During the shaping phase millet seeds were placed in front of the centre slit and mice used both paws to reach for them. Shaping was considered finished when 20 reach attempts were achieved within 20 min, and the mouse showed >70% limb preference. Training started the day after shaping, and each training day consisted of one session of 30 trials with preferred limb or 20 min (whichever occurred first). Seeds were presented individually in front of the slit on the side of preferred limb. Occasionally a mouse used the non-preferred limb; however, because of the difficulties presented by reaching angle, such reaches usually were unsuccessful. Mice displayed three reach attempt types: fail, drop and success (Supplementary Movie 1). A 'fail' was scored as a reach in which the mouse failed to touch the seed or knocked it away. A 'drop' was a reach in which the mouse retrieved the seed, but dropped it before putting into its mouth. A 'success' was a reach in which the mouse successfully retrieved the seed and put it into its mouth. Success rates were calculated as the percentage of successful reaches over total reach attempts. About half of the mice in our experiments were right handed (55 right handed out of a total of 109 mice, 50.6%). All data collected from both left- and right-handed mice were pooled for analysis in this study. No significant difference was found in the reaching performance of left- and right-handed mice.

All our control mice were littermates that underwent the same food restriction. All mice were handled (that is, removed from their cages and placed temporarily in the training chamber into which some seeds were dropped) by the same experimenters. To ensure that the increase seen in spine dynamics was learning specific, three different controls were used in our study. The first control group was general controls comprising mice with neither training nor shaping, but with food restriction, food reward and handling. The second was shaping controls in which mice received similar shaping as trained mice. During training, they were placed into the training chamber for 20 min daily, with ~15 seeds periodically dropped into the training chamber. This control group was used to determine whether the shaping period and/or experience of the training environment had any effect on spine dynamics. The third control group was activity controls in which mice were given similar shaping as trained mice. During training, mice were placed into the training chamber and trained to reach for a seed placed outside the slit for 20 min daily. However, the seed was placed out of reach, so that they could never obtain it and, therefore, did not learn skilful reaching movements (as shown by testing their performance occasionally). Thus, both trained mice and activity control mice experienced similar amounts of forelimb activity, but only trained mice developed the motor skill. The activity control was used to determine whether enhanced spine dynamics were caused by increased motor activity or were specific to motor skill learning. Our results indicate that there is no difference in the spine dynamics between the activity controls and general controls.

**Capellini handling task.** This task was similar to the vermicelli handling tasks previously described for rats[31]. Mice were food-restricted to maintain 90% of free feeding weight before training began. A daily training session consisted of 10 trials with uncooked capellini pasta pieces (2.5 cm), given one piece per trial. Mice learned to use coordinated forepaw movements to eat the pasta. The average consumption time for one piece of capellini pasta decreased from 3.44 ± 0.18 min on day 1 to 1.98 ± 0.29 min on day 4 (mean ± s.e.m., $P < 0.005$, 7 mice). There was no significant behavioural difference in the capellini handling task between naive adults and adults pre-trained in the reaching task in adolescence.

***In vivo* imaging of superficial dendrites.** The procedure for transcranial two-photon imaging has been described previously[8]. Mice aged 1–6 months were anaesthetized with an intraperitoneal injection (5.0 ml per kg body weight) of 17 mg ml$^{-1}$ ketamine and 1.7 mg ml$^{-1}$ xylazine in 0.9% NaCl. The skull was exposed with a midline scalp incision and imaged regions were located based on stereotactic coordinates. A small region of skull (~300 μm in diameter) was manually thinned down to ~20 μm in thickness using both a high-speed drill and a microknife. To reduce respiration-induced movements, the skull was glued to a 400-μm-thick stainless steel plate with a central opening for skull access. The plate was screwed to two lateral bars located on either side of the head and fixed to a metal base. The brain of the mouse was then imaged through the thinned skull using a Prairie Ultima IV multi-photon microscope with a Ti-sapphire laser tuned to the excitation wavelength for YFP (925 nm). Stacks of image planes were acquired with a step size of 0.70 μm using a water-immersion objective (×60, NA 1.1 infrared Olympus objective) at a zoom of 3.0. For relocation of the same dendrites at subsequent imaging times, an image stack containing the dendritic structures of interest was taken without zoom with a step size of 2.0 μm and the surrounding blood vessels were imaged with a CCD camera. The patterns of blood vessels and neuronal processes in this low-resolution image stack were used for relocating the same dendrites at each subsequent imaging session (Supplementary Fig. 1). After imaging, the plate was detached from the skull, the scalp sutured, and the animal was returned to its home cage until the next imaging session.

**Spine and filopodium identification.** All analysis of spine dynamics was done manually using ImageJ software, blind with regard to experimental conditions. Briefly, the same dendritic segments (~5–20 μm in length) were identified from three-dimensional image stacks selected from all views having high image quality (signal-to-background-noise ratio >4-fold). Individual dendritic protrusions were tracked manually along dendrites. Three-dimensional stacks, instead of two-dimensional projections, were used for analysis to ensure that tissue movements and rotation between imaging intervals did not influence identification of dendritic protrusions. The number and location of dendritic protrusions (protrusion length >1/3 dendritic shaft diameter) were identified in each view. Filopodia were identified as long thin structures with head diameter/neck diameter <1.2 and length/neck diameter >3. The remaining protrusions were classified as spines.

**Analysis of spine and filopodial dynamics.** Notations of the formation and elimination of spines and filopodia were based on comparison of the images collected at two different time points. Spines or filopodia were considered the same between two views if they were within 0.7 μm of their expected positions, based on their spatial relationship to adjacent landmarks and/or their position relative to immediately adjacent spines. A stable spine is a spine that was present in both images. An eliminated spine is a spine that appeared in the initial image, but not the second image. A newly formed spine is a spine that appeared in the second image, but was absent from the initial image. Percentages of stable, eliminated and formed spines were all normalized to the initial image. Percentage changes in the total spine number over a given interval were relative to the first view and calculated as percentage of formation minus percentage of elimination measured over that interval. Data on spine dynamics are presented as mean ± s.d..

**Image processing and presentation.** Two-dimensional projections of three-dimensional image stacks containing in-focus dendritic segments of interest were used for all figures. We chose very sparsely labelled regions as examples and maximum projections were made from images from 2–4 focus planes. There were normally few crossing structures in the projected images from such a shallow stack, and the presented branches could be clearly isolated. Finally, images were thresholded, Gaussian filtered and contrasted for presentation.

**Mapping of motor cortex by intracortical microstimulation.** This method was adapted from those used in rat experiments[21]. Mice were anaesthetized with an initial cocktail of ketamine (150 mg kg$^{-1}$, intraperitoneal) and xylazine (10 mg kg$^{-1}$, intraperitoneal) and supplemented with additional ketamine and isofluorane (0.5–1% in oxygen) as necessary. The mouse was placed into a mouse stereotaxic frame (Stoelting), lidocaine (2 mg kg$^{-1}$, subcutaneous) was injected into the scalp, and a midline incision was made. The cisterna magna was drained to prevent cortical swelling and the skull and dura overlying the motor cortex were removed. The craniotomy was then filled with warm (37 °C) silicone oil to prevent drying. A picture of the cortical surface was taken and overlaid with a 250 μm square grid in Canvas software.

Intracortical penetrations of a glass microelectrode (diameter of 20-25 μm) with a platinum wire were made at 250 μm intervals in a systematic order throughout the cortex at a depth of 790–800 μm (corresponding to deep layer V/shallow layer VI) with a hydraulic micropositioner until the entire extent of the forelimb representation was resolved. A 40-ms train of 13 200-μs monophasic cathodal pulses was delivered at 350 Hz from an electrically isolated, constant current stimulator at a rate of 1 Hz stimulation and current was increased to a maximum of 60 μA until a visible movement was evoked. If a movement was evoked at or below 60 μA, the threshold current was determined by gradually decreasing the stimulation until the movement stopped. The lowest current that evoked a movement was taken as the threshold current. If no movement was seen at 60 μA, the site was considered non-responsive. In cases where stimulation evoked more than one movement, the site was considered responsive to the movement that was determined to have the lowest threshold. To verify that the stimulation position was located within layer V, we injected DiI in seven mice at the end of the experiments and found that all injections left deposits extending through mid layer V to mid layer VI. In addition, penetrating electrode tracts could be observed in Nissl-stained coronal sections in most mice. Most (81.3 ± 4.7%) of these tracts terminated in layer V at a measured depth of 782 ± 137 μm, with the remainder terminated in upper layer VI.

31. Allred, R. P. *et al.* The vermicelli handling test: a simple quantitative measure of dexterous forepaw function in rats. *J. Neurosci. Methods* **170**, 229–244 (2008).

Vol 462|17 December 2009|doi:10.1038/nature08577

LETTERS

# Stably maintained dendritic spines are associated with lifelong memories

Guang Yang[1], Feng Pan[1] & Wen-Biao Gan[1]

**Changes in synaptic connections are considered essential for learning and memory formation[1–6]. However, it is unknown how neural circuits undergo continuous synaptic changes during learning while maintaining lifelong memories. Here we show, by following postsynaptic dendritic spines over time in the mouse cortex[7,8], that learning and novel sensory experience lead to spine formation and elimination by a protracted process. The extent of spine remodelling correlates with behavioural improvement after learning, suggesting a crucial role of synaptic structural plasticity in memory formation. Importantly, a small fraction of new spines induced by novel experience, together with most spines formed early during development and surviving experience-dependent elimination, are preserved and provide a structural basis for memory retention throughout the entire life of an animal. These studies indicate that learning and daily sensory experience leave minute but permanent marks on cortical connections and suggest that lifelong memories are stored in largely stably connected synaptic networks.**

One remarkable feature of the mammalian brain is its capacity to integrate new information throughout life while stably maintaining memories. Coincident with these two seemingly mutually exclusive attributes of the brain are plasticity and stability of synaptic connections[1–11]. It is well-established that the strength and number of synaptic connections can undergo rapid and extensive changes after sensory alterations and learning throughout life[1,2,4,6,9,12–19]. On the other hand, recent studies have shown that dendritic spines, the postsynaptic sites of excitatory synapses, are remarkably stable in adult life[7–9]. Therefore, synaptic connections are not only capable of undergoing rapid changes in response to new experience but also can serve as substrates for long-term information storage. However, it remains unknown how and to what degree synapses reorganize during learning and how such reorganization is transformed into lifelong memories.

To address these questions, we used transcranial two-photon microscopy to examine how fluorescently labelled dendritic spines of layer V pyramidal neurons in the mouse cortex are altered and maintained in response to skill learning or novel sensory experience[7,20–22]. We first examined spine dynamics in the primary motor cortex after motor skill learning on an accelerated rotarod[20,23] (see Methods). In this rotarod learning task, animals changed their gait pattern and learned specific movement strategies beyond simply running quickly[23]. In the forelimb area of the motor cortex, rotarod training over 2 days leads to a significant increase (~5–7%) in spine formation in both young (1 month of age) and adult (>4 months) mice ($P < 0.001$; Fig. 1a–f and Supplementary Table). The increased spine formation was not observed in mice subjected to running similar distances on a slowly rotating rotarod and was region specific, occurring in the forelimb motor cortex but not in the barrel cortex (Fig. 1f). Notably, after being trained for 2 days, spine formation over the next 2 days remained significantly higher if mice were trained with a different type of motor task (reverse

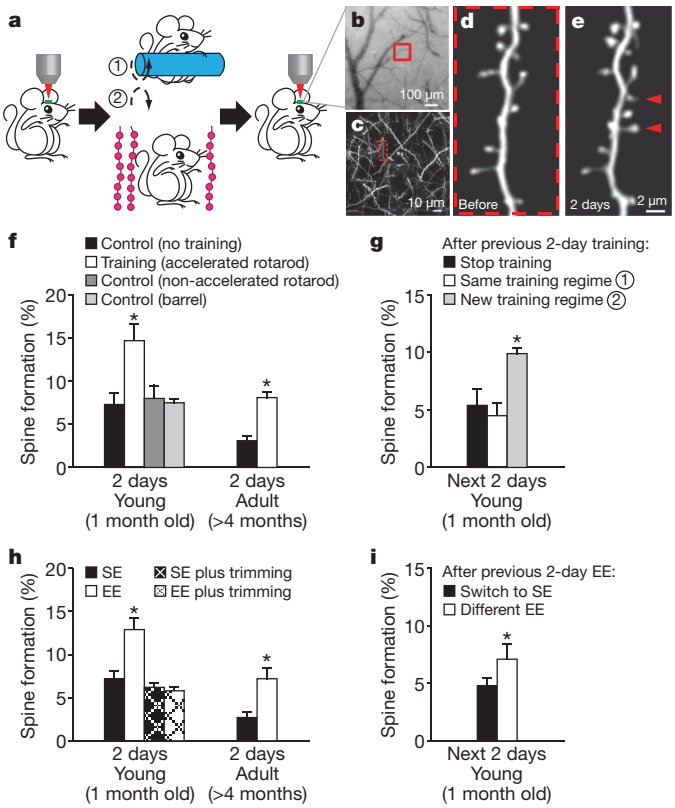



**Figure 1 | Motor learning and novel sensory experience promote rapid dendritic spine formation.** **a**, Transcranial two-photon imaging of spines before and after rotarod training or sensory enrichment. **b**, CCD camera view of the vasculature of the motor cortex. **c**, Two-photon image of apical dendrites from the boxed region in **b**. A higher magnification view of a dendritic segment in **c** is shown in **d**. **d**, **e**, Repeated imaging of a dendritic branch before (**d**) and after rotarod training (**e**). Arrowheads indicate new spines formed over 2 days. **f**, The percentage of new spines formed within 2 days in the motor cortex was significantly higher in young or adult mice after training as compared with controls with no training or running on a non-accelerated rotarod. No increase in spine formation was found in the barrel cortex after training. **g**, After previous 2-day training, only a new training regime (reverse running) caused a significant increase in spine formation. **h**, EE increased spine formation over 2 days in the barrel cortex in both young and adult animals. No significant increase in spine formation was found under EE when the whiskers were trimmed. **i**, After previous 2-day EE, animals switched to a different EE showed a higher rate of spine formation than those returned to SE. Data are presented as mean ± s.d. *$P < 0.005$. See Supplementary Table for the number of animals in each group.

[1]Molecular Neurobiology Program, The Helen and Martin Kimmel Center for Biology and Medicine at the Skirball Institute of Biomolecular Medicine, Department of Physiology and Neuroscience, New York University School of Medicine, New York, New York 10016, USA.

running) than if mice were subjected to the same type of training or no training ($P < 0.005$; Fig. 1a, g). Thus, motor learning experience, not just physical exercise, induces rapid spine formation within 2 days in the primary motor cortex.

To further understand experience-dependent spine plasticity, we examined the impact of novel sensory experience on spine formation in the barrel cortex, the primary somatosensory area for whisker sensation, by switching animals from a standard housing environment (SE) to an enriched environment (EE) (see Methods). When either young or adult mice were switched from SE to EE, spine formation over 1–2 days was significantly (~5%) higher than that under SE (Fig. 1a, h; $P < 0.001$; Supplementary Fig. 1). After being housed in an EE for 2 days, spine formation over the next 2 days remained significantly higher if mice were housed in a different EE than if mice were switched from EE to SE ($P < 0.005$; Fig. 1i). Notably, sensory deprivation by whisker trimming prevented the increase in spine formation associated with EE over 2 days ($P > 0.2$; Fig. 1h). Thus, novel sensory whisker experience, not simply the exploratory activity of the animals under EE, induces new spine formation in the barrel cortex. It is worth mentioning that regardless of animals' ages, neither EE nor motor learning increased the number of new dendritic filopodia, spine precursors[7,8,24], over 2 days (Supplementary Fig. 2). Together, these findings indicate that at different stages of animals' lives, learning and novel sensory experience induce rapid and extensive spine formation in functionally relevant cortical regions.

To gain insights into the functional significance of new spines, we examined the maintenance of new spines under various conditions (with or without skill learning, housed under EE or SE). We found that regardless of the animals' ages or conditions, a small fraction of new spines formed over 2 days remained over the next 2 weeks whereas most new spines (>75%) were eliminated (Fig. 2a–c and Supplementary Fig. 3). Interestingly, a significantly larger fraction of new spines lasted over 2 weeks when the mice were trained for 4–14 consecutive days than when they were trained for only 2 days ($P < 0.05$; Fig. 2b). Similarly, a larger fraction of new spines remained if mice continued to stay in the EE than if they were switched from EE to SE or stayed under EE but with their whiskers trimmed (Fig. 2c). Thus, although new spines are rapidly induced by novel experience (Fig. 1f, h), only a small fraction of them are maintained over weeks by a protracted process facilitated by persistent experience.

Many lines of evidence suggest that functional reorganization of mammalian cortex associated with motor and sensory training consists of a fast phase (within an individual training session) and a slow phase (between training sessions)[22,25]. The improvement of performance between sessions reaches a plateau over days to weeks and can persist for months to years[20,22,23,25]. The survival of a fraction of new spines for weeks suggests that they may be important for slow-phase learning and memory retention. Indeed, we found that the proportion of new spines that were formed within the first 2 days and remained at day 7 highly correlated with the retention of learned motor skills, as quantified by the average running speed that mice mastered on an accelerated rotarod ($r = 0.93$; Fig. 2d and Supplementary Fig. 4). In contrast, the extent of new spines accumulated from the beginning of training until day 7 did not correlate with motor skill performance ($r = -0.13$; Fig. 2e), underscoring the importance of experience-specific spine formation rather than increased spine turnover in general. The strong correlation between maintained new spines and slow-phase learning suggests that new spines are important for the reorganization of cortical circuits that underlie new motor skills. Furthermore, because a fraction of new spines induced by novel sensory experience are maintained, they may be important for receptive field reorganization in barrel cortex and contribute to whisker-based decision making[2,26].

In addition to promoting synapse formation, experience plays an important role in eliminating excessive and imprecise synaptic connections formed early during development[3–6,9]. To understand experience-dependent synaptic remodelling further, we examined the elimination of early formed spines in young mice subjected to motor training or exposed to EE. We found that in 1-month-old mice, neither motor training nor novel sensory experience increased the elimination of existing spines or filopodia over 2 days in motor or barrel cortex ($P > 0.4$; Fig. 3a, b and Supplementary Fig. 2). However, a significant increase in spine elimination (~4.5%) was observed in

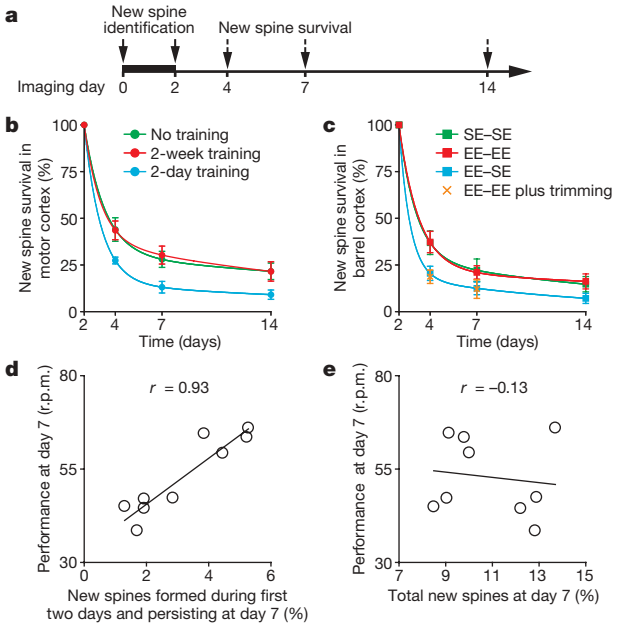



**Figure 2 | A fraction of newly formed spines persists over weeks and correlates with performance after learning.** **a**, New spines induced by novel experience were identified in the first 2 days and followed over time. **b**, **c**, The survival of new spines (mean ± s.d.) over time under various conditions. A significantly larger fraction of new spines remained in mice trained repeatedly or housed under EE continuously. The lines represent two exponential fittings ($r^2 = 1$). **d**, **e**, An animal's performance at day 7 strongly correlated with new spines formed during the first 2-day training and persisting at day 7 (**d**), but did not correlate with the total new spines accumulated from day 0 to 7 (**e**). Each circle represents an individual animal. The linear regression lines and correlation coefficients ($r$) are shown. See Supplementary Table for the number of animals in each group.

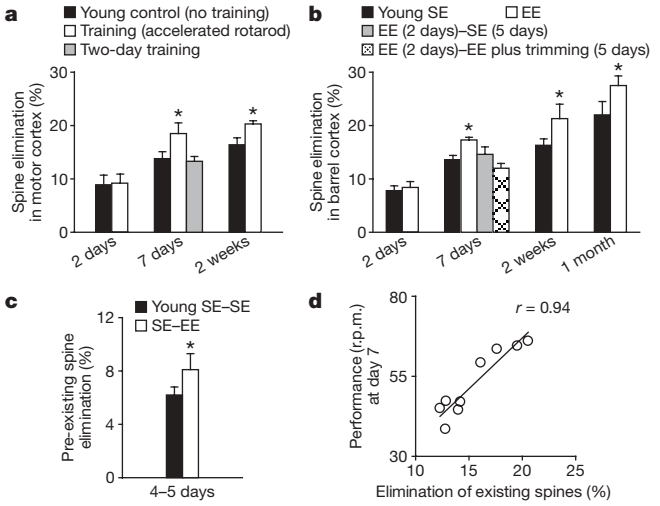



**Figure 3 | Novel experience promotes spine elimination.** **a**, **b**, Percentage of spines eliminated (mean ± s.d.) in young animals under various conditions. Rotarod training (**a**) or EE (**b**) for at least 7 days increased the elimination of existing spines ($P < 0.05$). **c**, EE increased the elimination of spines that existed for more than 2 days before EE exposure ($P < 0.05$). **d**, The elimination of existing spines over 7 days strongly correlated with an animal's performance on day 7 ($r = 0.94$). Each circle represents an individual animal.

motor cortex when mice were subjected to training for 7–14 days ($P < 0.05$; Fig. 3a). Similarly, more spines were eliminated in barrel cortex if mice continued to stay in EE for 7–30 days than if they were switched from EE to SE or stayed under EE but with their whiskers trimmed ($P < 0.05$; Fig. 3b). Furthermore, we found that the elimination of spines that have existed for at least 2 days was increased by new experience over 4–5 days ($P < 0.05$; Fig. 3c). Because the spines in this pool have likely all made synaptic contacts with axonal terminals[15,24], these results suggest that new experience leads to the pruning of existing synapses and could cause significant functional changes in cortical circuits. Indeed, we found that 1 week after motor training, motor performance strongly correlated with the degree of spine elimination ($r = 0.94$; Fig. 3d). Thus, motor learning and novel sensory experience involve not only new spine formation but also permanent removal of connections established early in life (Supplementary Fig. 5).

Although the above findings are consistent with the general notion that structural synaptic plasticity is critical for learning and memory, they raise a fundamental issue about how ongoing experience-induced synaptic reorganization can be reconciled with the stability needed to support lifelong memories. To address this issue, we first examined whether new spines could be maintained over a lifetime. If a significant number of new spines could last throughout an animal's lifespan, they could directly contribute to permanent memory storage. Otherwise, lifelong memory storage cannot rely on these new spines and may involve continuous rewiring of synaptic networks.

To distinguish between these possibilities, we examined the survival of new spines over many months in motor and barrel cortices. We found that ~4–5% of new spines formed over 2 days persisted for at least 3 months in motor cortex (2 of 42 spines formed after 2-day training) and for at least 5 months in barrel cortex (2 of 50 new spines). Thus, a tiny fraction of daily formed new spines (~0.2% of the total spines) could persist for 3–5 months. Because it is difficult to measure directly and accurately a small fraction of new spines surviving over many months, we estimated long-term survival of new spines based on the fact that the accumulation of new spines depends on the formation rate of new spines and their survival fraction (Fig. 4a and Supplementary Information 1). Because the rate of spine formation is relatively constant throughout adult life (Supplementary Fig. 6) and the survival fraction of new spines is comparable under a constant environment (Figs 2b, c and 5), we found that our direct measurement of new spine accumulation over time in barrel cortex can be best fitted by three exponential components with time constants of ~1.5 days, ~1–2 months and ~73–80 months, respectively (Fig. 4a and Supplementary Information 1). The first two exponential components suggest that most daily formed spines have an average lifetime of ~1.5 days and a small fraction have an average lifetime of ~1–2 months. Importantly, the third component suggests that ~0.8% of daily formed new spines have an average lifetime of ~80 months under SE and ~73 months under EE (Fig. 4a and Supplementary Information 1). Because the degree of spine formation and the survival of new spines are comparable between motor and barrel cortices, a similar degree of daily generated new spines in motor cortex are also expected to last over the entire life of an animal.

Based on the survival function of new spines and ~5–7% spine formation over 2 days under EE or motor learning conditions (Fig. 1), we estimated that the number of new spines formed over 2 days and persisting at the end of life would be ~0.04% of the total spines in motor or barrel cortex (assuming the mouse lifespan is ~36 months; Supplementary Information 2). Given the large quantity of spines in the mouse cortex, the number of learning-induced and subsequently maintained new spines could be $\sim 2 \times 10^6$, sufficiently many to have a significant and lifelong impact on neural network functions and an animal's behaviour[27,28] (Supplementary Information 2).

Although a fraction of daily generated spines persist and could directly contribute to lifelong memory storage, it is important to

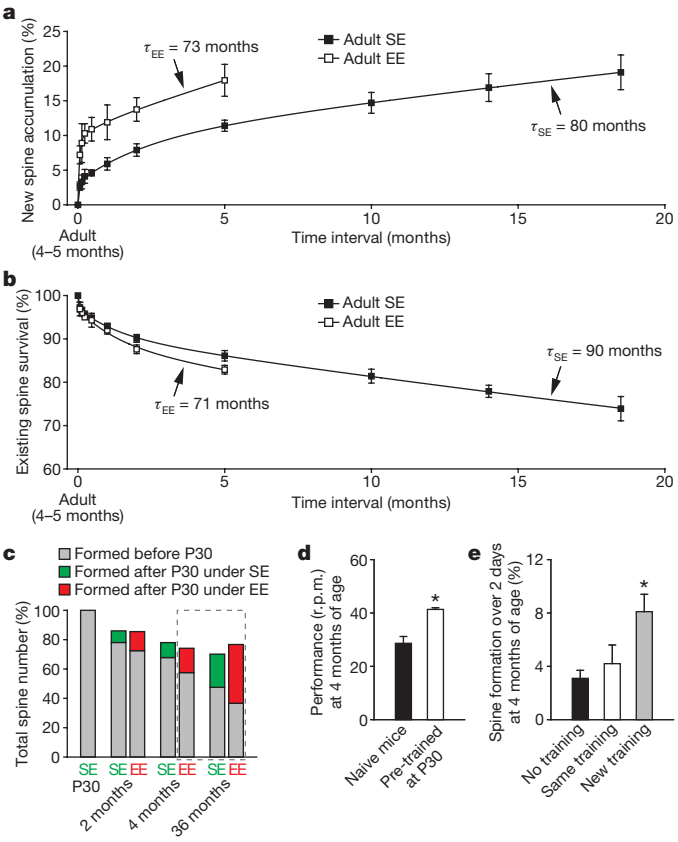



**Figure 4 | Maintenance of daily formed new spines and spines formed during early development throughout life.** **a**, New spine accumulation over time under SE and EE. Three exponential fits show that ~0.8% of daily formed new spines decay with a time constant of 80 months under SE and 73 months under EE. **b**, The percentage of adult spines remaining over time under SE and EE. Three exponential fits show that ~90% of adult spines have an average lifetime of 90 months under SE and 71 months under EE. **c**, A large fraction of spines formed before P30 persisted throughout life under SE or EE. The projections based on **a** and **b** are shown in the dashed frame. **d**, Mice previously trained at P30 for 7 days showed better performance (mean ± s.e.m.) when assessed at 4 months of age than naive mice ($P < 0.01$). **e**, Only a new training regime (reverse running) caused an increase in spine formation in previously trained animals. Spine data are presented as mean ± s.d.

note that they represent a minute portion (~0.04%) of the total spine population at the end of an animal's life and likely have their impact on the animal's behaviour in the context of existing circuitry rather than acting alone. Because the pruning of existing spines is an important aspect of learning (Fig. 3), this raises the question of whether early formed spines would persist throughout adult life. If a fraction of early formed spines were maintained over a lifetime, they may serve as substrates for preserving basic cortical functions and early memories. Otherwise, the physical substrates of early memories would have to be re-established in synaptic networks that are formed later in adulthood.

To address this question, we measured the survival of existing spines over many months in barrel cortex under SE and EE. We found that in 4-month-old adult mice, ~86% and ~83% of existing spines are maintained over a period of 5 months under SE and EE, respectively (Fig. 4b). Based on the survival of existing spines over 5–18.5 months, we estimated that ~90% adult spines have an average lifetime of ~90 months under SE and ~71 months under EE (Fig. 4b and Supplementary Information 3). Furthermore, we found that ~78% and ~73% of existing spines are maintained from postnatal day 30 (P30) to 2 months of age under SE and EE, respectively (Fig. 3b). Assuming a lifespan of 36 months, ~48% (under SE) and ~37% (under EE) of spines existing at P30 would remain at the end of life (Fig. 4c). Thus,

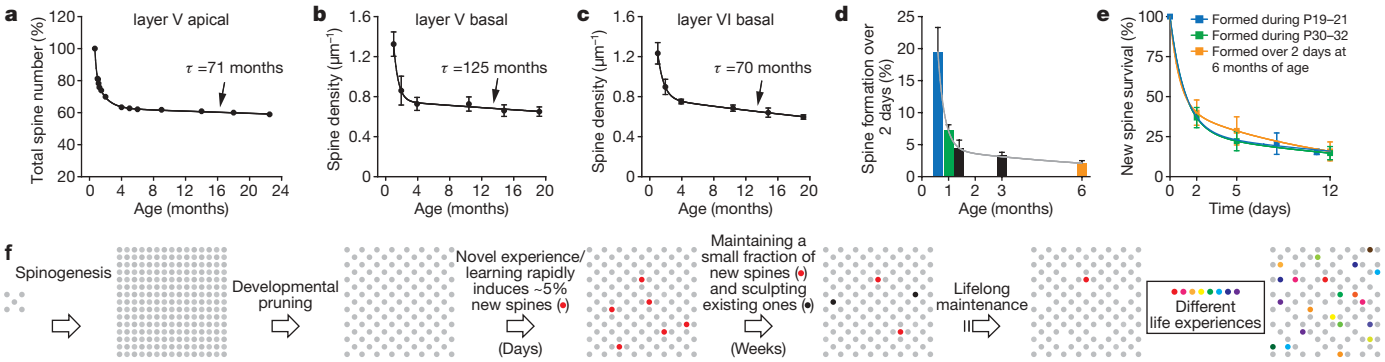



**Figure 5 | Spine maintenance in different cell types and cortical layers.** **a–c**, Age-dependent change in spine number is remarkably similar across different cell types/cortical layers in barrel cortex and contains information on spine dynamics. Total spine number (percentage of P19) of layer V pyramidal cell apical dendrites (**a**) was measured through *in vivo* imaging. Spine densities (mean ± s.e.m.) of layer V and layer VI pyramidal cell basal dendrites (**b, c**) were measured on dendritic segments located 50–100 μm from the soma in fixed brain slices. **d**, Spine formation rate declined rapidly from P19 to P30 and remained low thereafter. **e**, Regardless of animals' ages (P19, P30, 6 months), a fraction of new spines formed over 2 days were maintained over a similar protracted process. **f**, Schematic summary of spine remodelling and maintenance throughout life. Spines are rapidly formed after birth, undergo experience-dependent pruning during postnatal development and remain largely stable in adulthood. Learning or novel sensory experience induces rapid formation of new spines (~5% of total spines) within 1–2 days. Only a tiny fraction of new spines (~0.04% of total spines) survive the first few weeks in synaptic circuits and are stably maintained later in life. Novel experience also results in the pruning of a small fraction of existing spines formed early during development. New stable spines induced by novel experience, together with existing spines formed during early development and surviving experience-dependent pruning, provide an integrated and stable structural basis for lifelong memory storage, despite ongoing plasticity in synaptic networks.

regardless of housing environments, a large fraction of spines that are formed before P30 in barrel cortex would persist throughout life. Because motor learning and novel sensory experience lead to a similar degree of spine remodelling in either young or adult mice (Figs 1–3 and Supplementary Table), a large fraction of early formed spines are also expected to be stably maintained in the motor cortex. Together, these results suggest that spines formed early during development and surviving experience-dependent elimination could provide a scaffold for basic cortical function and lifelong memory storage.

By examining how spines reorganize and maintain in response to novel experiences (Figs 1–4), our studies have revealed the existence of two populations of stable spines in synaptic circuits. One population constitutes new spines specifically induced by novel experience and maintained later in life. The other population comes from a large spine pool formed during early postnatal development, pruned by developmental experiences and surviving throughout adulthood. Because spines in both populations have an average lifetime between 70 and 90 months (Fig. 4a, b), ~60–70% of them could persist over an animal's life and directly support lifelong memories in synaptic circuits.

One prediction from such a synaptic model of memory storage is that information should still be maintained even though ~30–40% of synapses in the circuitry are lost. To test this experimentally, we trained animals on the rotarod task from P30 to P37 and tested their performance at 4 months of age, when ~30% of spines that existed at P30 were eliminated in barrel and motor cortices (Fig. 4c and Supplementary Table). We found that animals previously trained at P30 could still maintain their learned motor skills when tested again at 4 months of age (Fig. 4d). Notably, the same training regime did not result in a significant increase in spine formation over 2 days in these previously trained mice ($P > 0.2$), whereas a different training regime did ($P < 0.02$) (Fig. 4e). These findings are consistent with the above synaptic model of memory storage, suggesting that dynamic (~30% spine loss) but largely stable circuits could maintain previously acquired skills.

By studying spine dynamics of layer V pyramidal cell apical dendrites, our results suggest that spine maintenance is a fundamental feature of neural circuits important for memory storage. However, it remains unclear whether the same rule regulating spine dynamics on layer V apical dendrites applies to spines in other cell types or cortical layers or regions. As shown below, by analysing age-dependent developmental profiles of spine number, we found evidence that stably maintaining a fraction of new spines and spines formed early in life is likely to be a general rule for lifelong information storage in the cortex.

Many lines of evidence indicate that developmental change in synapse number is remarkably similar across different cortical layers and regions in a variety of species[7,8,29,30]. We found that in the dendrites of layer V and VI pyramidal neurons in mouse barrel cortex, the number of spines rose rapidly after birth, underwent a substantial net loss during late postnatal life and declined slowly throughout adulthood (Fig. 5a–c). Importantly, in the apical dendrites of layer V pyramidal cells, we found that the substantial net loss of spines during postnatal development was due to a combination of two factors: (1) a tremendous burst in spine formation early in life was followed by a rapid decline in spine formation from P19 to P30 (Fig. 5d); and (2) regardless of developmental stages, only a small fraction of newly formed spines were maintained by a similar prolonged process (Fig. 5e and Supplementary Fig. 7). Specifically, a substantial net loss of spines occurred during the late postnatal period because early formed spines (before P19) continued to be eliminated from P19 to P60 at a rate higher than that of new spine addition. The remarkably similar patterns of developmental spine loss in different cortical layers and species suggest that both a rapid decline in spine formation and maintenance of a fraction of new spines by a prolonged process are general rules in the development of the mammalian cortex (Supplementary Information 4).

If stably maintaining a fraction of new spines by a prolonged process is a common rule, do most adult spines in other cells and layers persist as those of layer V pyramidal cell apical dendrites? Assuming that spine formation is constant throughout adulthood and all spines that survive a prolonged process have the same average lifetime, $\tau$, the total number of adult spines would change according to the equation $A + Be^{-t/\tau}$ (Supplementary Information 5). Based on the gradual decline in spine number of apical dendrites of layer V pyramidal cells (Fig. 5a), we estimated that the average lifetime of adult spines is ~71 months. This number is highly comparable to the average lifetime of new stable spines (Fig. 4a; 73–80 months) or existing spines (Fig. 4b; 71–90 months) that we measured with the *in vivo* imaging approach. Furthermore, based on the age-dependent decline in spine density of basal dendrites of layer V and VI pyramidal cells (Fig. 5b, c), we estimated that the average lifetime of adult spines on these dendrites is ~70–125 months. Together, these projections

suggest that (1) developmental profiles of spine number contain important information on spine dynamics, and (2) most adult spines in other cell types and cortical layers could be stably maintained and serve as substrates for long-term information storage.

Determining how long-lasting memories are stored in neuronal circuits remains a great challenge. Because synapses undergo rapid changes in response to environmental perturbations, it is unknown how dynamic synaptic circuits maintain indelible memories. Here we show that, despite ongoing circuit plasticity, two populations of stable spines are important for maintaining lifelong memories. Specifically, our findings suggest that a minute fraction of new spines (~0.04% of total spines) induced by novel experience, together with spines formed early during development and remaining after experience-dependent pruning, represent a unique and stable physical entity for lifelong memory storage (Fig. 5f and Supplementary Discussion). The fact that most spines in such an entity persist underscores the fundamental importance of stably connected synaptic circuits in lifelong memory storage.

## METHODS SUMMARY

Mice expressing YFP (H-line) were used in all the experiments. Sensory enrichment was conducted by placing mice in standard mouse cages containing strings of beads whose positions were changed daily. Motor training was performed by placing mice on an accelerated motorized rod. The rotation speed was recorded when the animal could not keep up with the rotating-rod and fell. The performance was measured as the average speed animals achieved during the 20-trial training session per day. The procedure of *in vivo* transcranial two-photon imaging, spine density measurement and data quantification was described previously[7,8]. *P* values were calculated using Student's *t*-test.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 14 August; accepted 12 October 2009.**
**Published online 29 November 2009.**

1. Bailey, C. H. & Kandel, E. R. Structural changes accompanying memory storage. *Annu. Rev. Physiol.* **55,** 397–426 (1993).
2. Buonomano, D. V. & Merzenich, M. M. Cortical plasticity: from synapses to maps. *Annu. Rev. Neurosci.* **21,** 149–186 (1998).
3. Changeux, J. P. & Danchin, A. Selective stabilisation of developing synapses as a mechanism for the specification of neuronal networks. *Nature* **264,** 705–712 (1976).
4. Hubel, D. H., Wiesel, T. N. & LeVay, S. Plasticity of ocular dominance columns in monkey striate cortex. *Phil. Trans. R. Soc. Lond. B* **278,** 377–409 (1977).
5. Lichtman, J. W. & Colman, H. Synapse elimination and indelible memory. *Neuron* **25,** 269–278 (2000).
6. Shatz, C. J. & Stryker, M. P. Ocular dominance in layer IV of the cat's visual cortex and the effects of monocular deprivation. *J. Physiol. (Lond.)* **281,** 267–283 (1978).
7. Grutzendler, J., Kasthuri, N. & Gan, W. B. Long-term dendritic spine stability in the adult cortex. *Nature* **420,** 812–816 (2002).
8. Zuo, Y., Lin, A., Chang, P. & Gan, W. B. Development of long-term dendritic spine stability in diverse regions of cerebral cortex. *Neuron* **46,** 181–189 (2005).
9. Zuo, Y., Yang, G., Kwon, E. & Gan, W. B. Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex. *Nature* **436,** 261–265 (2005).
10. Purves, D. & Hadley, R. D. Changes in the dendritic branching of adult mammalian neurones revealed by repeated imaging *in situ*. *Nature* **315,** 404–406 (1985).
11. Trachtenberg, J. T. *et al.* Long-term *in vivo* imaging of experience-dependent synaptic plasticity in adult cortex. *Nature* **420,** 788–794 (2002).
12. Darian-Smith, C. & Gilbert, C. D. Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature* **368,** 737–740 (1994).
13. Sin, W. C., Haas, K., Ruthazer, E. S. & Cline, H. T. Dendrite growth increased by visual activity requires NMDA receptor and Rho GTPases. *Nature* **419,** 475–480 (2002).
14. Kleim, J. A., Vij, K., Ballard, D. H. & Greenough, W. T. Learning-dependent synaptic modifications in the cerebellar cortex of the adult rat persist for at least four weeks. *J. Neurosci.* **17,** 717–721 (1997).
15. Hofer, S. B., Mrsic-Flogel, T. D., Bonhoeffer, T. & Hubener, M. Experience leaves a lasting structural trace in cortical circuits. *Nature* **457,** 313–317 (2009).
16. Holtmaat, A., Wilbrecht, L., Knott, G. W., Welker, E. & Svoboda, K. Experience-dependent and cell-type-specific spine growth in the neocortex. *Nature* **441,** 979–983 (2006).
17. Dunaevsky, A., Tashiro, A., Majewska, A., Mason, C. & Yuste, R. Developmental regulation of spine motility in the mammalian central nervous system. *Proc. Natl Acad. Sci. USA* **96,** 13438–13443 (1999).
18. Matsuzaki, M., Honkura, N., Ellis-Davies, G. C. & Kasai, H. M. Structural basis of long-term potentiation in single dendritic spines. *Nature* **429,** 761–766 (2004).
19. Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R. & Muller, D. LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402,** 421–425 (1999).
20. Costa, R. M., Cohen, D. & Nicolelis, M. A. Differential corticostriatal plasticity during fast and slow motor skill learning in mice. *Curr. Biol.* **14,** 1124–1134 (2004).
21. Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning fluorescence microscopy. *Science* **248,** 73–76 (1990).
22. Karni, A. *et al.* The acquisition of skilled motor performance: fast and slow experience-driven changes in primary motor cortex. *Proc. Natl Acad. Sci. USA* **95,** 861–868 (1998).
23. Buitrago, M. M., Schulz, J. B., Dichgans, J. & Luft, A. R. Short and long-term motor skill learning in an accelerated rotarod training paradigm. *Neurobiol. Learn. Mem.* **81,** 211–216 (2004).
24. Ziv, N. E. & Smith, S. J. Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* **17,** 91–102 (1996).
25. Karni, A. & Sagi, D. The time course of learning a visual skill. *Nature* **365,** 250–252 (1993).
26. Celikel, T. & Sakmann, B. Sensory integration across space and in time for decision making in the somatosensory system of rodents. *Proc. Natl Acad. Sci. USA* **104,** 1395–1400 (2007).
27. Arenz, A., Silver, R. A., Schaefer, A. T. & Margrie, T. W. The contribution of single synapses to sensory representation *in vivo*. *Science* **321,** 977–980 (2008).
28. Houweling, A. R. & Brecht, M. Behavioural report of single neuron stimulation in somatosensory cortex. *Nature* **451,** 65–68 (2008).
29. Huttenlocher, P. R. & Dabholkar, A. S. Regional differences in synaptogenesis in human cerebral cortex. *J. Comp. Neurol.* **387,** 167–178 (1997).
30. Rakic, P., Bourgeois, J. P., Eckenhoff, M. F., Zecevic, N. & Goldman-Rakic, P. S. Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science* **232,** 232–235 (1986).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** This work was supported by National Institutes of Health R01 NS047325 and a Dart Foundation Fellowship to W.-B.G. and by an Ellison/AFAR Postdoctoral Fellowship to G.Y. We thank members of the Gan laboratory for their comments.

**Author Contributions** G.Y. and W.-B.G. conceived the experiments. G.Y. performed and analysed most experiments on motor cortex and all the experiments on barrel cortex. F.P. conducted and analysed some of the experiments on motor cortex. G.Y. and W.-B.G. performed the data fitting. W.-B.G. wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to W.-B.G. (gan@saturn.med.nyu.edu).

## METHODS

**Experimental animals.** Mice expressing YFP in layer V pyramidal neurons (H-line) were purchased from the Jackson Laboratory and group-housed in the Skirball animal facilities. All experiments were done in accordance with institutional guidelines.

**Sensory enrichment.** Sensory enrichment was conducted in standard mouse cages containing strings of beads hanging from the top of the cages (Supplementary Movies 1 and 2). The positions of bead strings were changed daily. Mice could move freely in these cages and had to navigate through the strings of beads to obtain food and water.

**Rotarod training procedure.** An EZRod system with a test chamber (44.5 cm × 14 cm × 51 cm, Accuscan Instruments) was used in this study. Animals were placed on the motorized rod (30 mm in diameter) in the chamber. The rotation speed gradually increased from 0 to 100 r.p.m. over the course of 3 min. The time latency and rotation speed were recorded when the animal was unable to keep up with the increasing speed and fell. Rotarod training/testing was performed in one 30-min session per day (20 trials in total). Performance was measured as the average speed animals achieved during the 20 trials. For control experiments, animals were either trialled 20 times by placing them on the still rod for 2 min, then dropped to the bottom of the chamber (no-training control), or by forcing them to run on the rod rotating at a constant speed of 15 r.p.m. (non-accelerated rotarod control, 60 min in total with a 20-s break every 5 min). A reverse running regime was introduced to provide pre-trained mice with a new motor learning experience. In this regime, animals were forced to run backwards on the rotating rod (speed increased gradually from 0 to 50 r.p.m. over 3 min) for 20 trials.

**Identification of the forelimb region of the motor cortex and the barrel cortex.** The location of imaging in the motor cortex is 1.3 mm anterior to the bregma and 1.2 mm lateral from the midline. In a previously published study[31], this region has been identified through microstimulation as the location of forelimb representations in the same mouse strain as we used in our study. We confirmed this region of forelimb representations by microstimulation in our own hands. In addition, the location of imaging in the barrel cortex is 1.1 mm posterior to the bregma and 3.4 mm lateral from the midline. We have previously confirmed this location is within the barrel cortex using cytochrome oxidase staining[9]. Because our imaging window was rather small (200 μm × 200 μm), we chose to use stereotaxic coordinates of previously mapped forelimb and barrel regions as the guide to study spine dynamics in motor and barrel cortices.

***In vivo* transcranial two-photon imaging.** The degree of spine formation and elimination was obtained from longitudinal studies by imaging the mouse cortex through a thinned-skull window. Because thinning the skull to ~20 μm at each imaging session without damaging the cortex becomes difficult after several chronic imaging sessions, we designed our experiments such that the same animals were imaged no more than four times. For the measurement of new spine survival in Fig. 2b, c, most but not all of the data came from chronic imaging of the same mice. For the measurement of new spine accumulation and existing spine survival in Fig. 4a, b, a total of 57 animals were used (most of them were imaged twice, eight of them were imaged three or four times).

The surgery and imaging procedures are described below.

1. Anaesthetize the mouse with an intraperitoneal injection of ketamine/xylazine mix (20 mg ml$^{-1}$ ketamine, 3 mg ml$^{-1}$ xylazine in saline, 5–6 μl g$^{-1}$ body weight).
2. Carefully shave the hair of the scalp with a double-edged razor blade. Make a midline incision of the scalp with sterile surgical scissors. The incision should extend from the middle of the ears to the frontal area.
3. Remove the periosteum tissue with a microsurgical blade. The brain area to be imaged was localized based on the stereotactic coordinates and marked with a fine marker.
4. Place a small amount of glue around the edges of the internal opening of the skull holding plate and press it against the skull for a few seconds. Make sure that the area to be imaged is exposed in the centre of the internal opening of the skull holder.
5. Wait approximately 5 min until the plate is stably glued to the skull and then place the mouse on a cotton pad on top of an optional heating pad. Attach the skull holder to two metal cubes adhered to a large plate for immobilizing the holder. Wash away unpolymerized glue with artificial cerebrospinal fluid (ACSF).
6. Use a high-speed micro-drill to thin a circular area of skull (typically ~0.5–1 mm in diameter) over the region of interest under a dissection microscope. Drilling should be done intermittently to avoid overheating. Replace ACSF periodically and wash away the bone debris.
7. The mouse skull consists of two thin layers of compact bone, sandwiching a thick layer of spongy bone. The spongy bone contains tiny cavities arranged in concentric circles and multiple canaliculi that carry blood vessels. Remove the external layer of the compact bone and most of the spongy bone with the drill. Some bleeding from the blood vessels running through the spongy bone may occur during the thinning process. This bleeding will usually stop spontaneously within a few minutes.
8. After removing most of the spongy bone, use a microsurgical blade to continue the thinning process until a very thin (~20 μm) and smooth preparation (~200 μm in diameter) is achieved.
9. Use a conventional epifluorescence microscope to check if dendrites and spines in the area of interest can be clearly visualized at this stage. The thickness of the skull can also be directly determined by visualization of the skull with a two-photon microscope.
10. A CCD (charge-coupled device) camera can be used to acquire a high-quality picture of the brain vasculature, which is used as a landmark for future relocation.
11. Carefully move the mouse to the two-photon microscope and select an area for two-photon imaging. The selected area is then carefully identified and marked in the CCD vasculature map.
12. Tune the two-photon microscope to the appropriate wavelength (920 nm for yellow fluorescent protein). Imaging is achieved by using ×60 water-immersion objectives with numerical aperture 1.1.
13. Obtain a low-magnification stack of fluorescently labelled neuronal processes at ×1 zoom, which serves as a more precise map for relocation of the same area at later time points in addition to the CCD image of brain vasculature. The stack is typically taken within ~200 μm below the pial surface. Additional higher magnification (×3 digital zoom) images can be taken by electronically moving the imaged area.
14. For re-imaging the same region, find the thinned region based on the brain vasculature map. Carefully remove the connective tissue that has re-grown on top of the thinned region using a microsurgical blade, and check the image quality with the two-photon microscope. The skull may need to be re-thinned.
15. Use a microsurgical blade to shave the skull carefully until a clear image can be obtained.
16. Find the imaged region under a fluorescence microscope. Align the region according to a ×1 zoom map under the two-photon microscope, then zoom in to ×3 to align it further.
17. After the image is precisely aligned with the first view, take images as previously described.

**Data analysis.** ImageJ software was used to analyse image stacks. The same dendritic segments were identified from three-dimensional stacks taken from different time points with high image quality (ratio of signal to background noise >4:1). The number and location of dendritic protrusions (protrusion length was more than one-third the dendritic shaft diameter) were identified in each view without previous knowledge of the animal's age, the interval between views or the order of the views. The total number of spines ($n$) was pooled from dendritic segments of different animals. Filopodia were identified as long, thin structures (generally larger than twice the average spine length, ratio of head diameter to neck diameter <1.2:1 and ratio of length to neck diameter >3:1). The remaining protrusions were classified as spines. No subtypes of spines were separated. Three-dimensional stacks were used to ensure that tissue movements and rotation between imaging intervals did not influence spine identification. Spines or filopodia were considered the same between views if their positions remained the same distance from relative adjacent landmarks. Spines were considered different if they were more than 0.7 μm away from their expected positions based on the first view.

Changes in cortical volume associated with motor skill learning and EE have been well documented previously[32–35]. It is important to note that because we measured spine dynamics on the same dendrites in the same animals over time, our measurements of spine elimination and formation were not sensitive to changes in cortical volume.


31. Li, C. X. & Waters, R. S. Organization of the mouse motor cortex studied by retrograde tracing and intracortical microstimulation (ICMS) mapping. *Can. J. Neurol. Sci.* **18,** 28–38 (1991).
32. Anderson, B. J., Eckburg, P. B. & Relucio, K. I. Alterations in the thickness of motor cortical subregions after motor-skill learning and exercise. *Learn. Mem.* **9,** 1–9 (2002).
33. Diamond, M. C. *et al.* Increases in cortical depth and glia numbers in rats subjected to enriched environment. *J. Comp. Neurol.* **128,** 117–126 (1966).
34. Grossman, A. W., Churchill, J. D., Bates, K. E., Kleim, J. A. & Greenough, W. T. A brain adaptation view of plasticity: is synaptic plasticity an overly limited concept? *Prog. Brain Res.* **138,** 91–108 (2002).
35. Kleim, J. A., Pipitone, M. A., Czerlanis, C. & Greenough, W. T. Structural stability within the lateral cerebellar nucleus of the rat following complex motor learning. *Neurobiol. Learn. Mem.* **69,** 290–306 (1998).

LETTERS

# Division and apoptosis of E2f-deficient retinal progenitors

Danian Chen[1], Marek Pacal[1], Pamela Wenzel[2], Paul S. Knoepfler[3], Gustavo Leone[2] & Rod Bremner[1]

**The activating E2f transcription factors (E2f1, E2f2 and E2f3) induce transcription and are widely viewed as essential positive cell cycle regulators. Indeed, they drive cells out of quiescence, and the 'cancer cell cycle' in *Rb1* null cells is E2f-dependent[1,2]. Absence of activating E2fs in flies or mammalian fibroblasts causes cell cycle arrest[3,4], but this block is alleviated by removing repressive E2f or the tumour suppressor p53, respectively[5–7]. Thus, whether activating E2fs are indispensable for normal division is an area of debate[1]. Activating E2fs are also well known pro-apoptotic factors, providing a defence against oncogenesis[8], yet E2f1 can limit irradiation-induced apoptosis[9,10]. In flies this occurs through repression of *hid* (also called *Wrinkled*; Smac/Diablo in mammals). However, in mammals the mechanism is unclear because Smac/Diablo is induced, not repressed, by E2f1[11], and in keratinocytes survival is promoted indirectly through induction of DNA repair targets[12]. Thus, a direct pro-survival function for E2f1–3 and/or its relevance beyond irradiation has not been established. To address E2f1–3 function in normal cells *in vivo* we focused on the mouse retina, which is a relatively simple central nervous system component that can be manipulated genetically without compromising viability and has provided considerable insight into development and cancer[2,13]. Here we show that unlike fibroblasts, *E2f1–3* null retinal progenitor cells or activated Müller glia can divide. We attribute this effect to functional interchangeability with Mycn. However, loss of activating E2fs caused downregulation of the p53 deacetylase Sirt1, p53 hyperacetylation and elevated apoptosis, establishing a novel E2f–Sirt1–p53 survival axis *in vivo*. Thus, activating E2fs are not universally required for normal mammalian cell division, but have an unexpected pro-survival role in development.**

During retinal development, a thin neuroblastic layer of progenitors undergoes extensive expansion from mouse embryonic day 11 (E11) to approximately postnatal day 8 (P8), generating post-mitotic differentiating cells that develop into the six major retinal neurons and Müller glia. To study E2f1–3 function during retinal progenitor expansion, floxed *E2f3* mice were crossed with $E2f1^{-/-}$, $E2f2^{-/-}$ and *α-Cre* (which deletes *E2f3* in peripheral retinal progenitors at E10[14]) mice. We then assessed Ki67 to mark all dividing cells, or phosphohistone H3 (PH3) and 5-bromodeoxyuridine (BrdU) incorporation to mark mitosis and S phase, respectively (Fig. 1 and Supplementary Fig. 1a). Some reduction was detected in the $E2f1^{-/-}$ retina, but the $E2f2^{-/-}E2f3^{-/-}$ double knockout retina was unaffected, and *E2f2* or *E2f3* loss did not affect $E2f1^{-/-}$ progenitors (Fig. 1a, b and data not shown). At least one activating E2f is essential for fibroblast division[4], so we expected a drastic phenotype in the $E2f1^{-/-}E2f2^{-/-}E2f3^{-/-}$ triple knockout retina where *E2f3* is deleted before progenitor expansion. Broad *Cre-IRES-GFP* transgene expression, as well as analysis of DNA, messenger RNA and protein at various time points, all demonstrated robust *E2f3* excision (Supplementary Fig. 1), yet remarkably we observed many triple knockout $Ki67^+$ and $PH3^+$ cells at both E14 and P0 (Fig. 1a, b). The effect was slightly more pronounced at P0 than E14, and is discussed later. At E14 and E17 BrdU labelling was weaker in triple knockout versus wild-type progenitors but many $BrdU^+$ cells were obvious at higher magnification, suggesting continued, albeit slower, division (Supplementary Fig. 1a). At E14 and P0 cell cycle distribution was the same in wild-type, $E2f1^{-/-}$ or triple knockout retinas (Supplementary Fig. 2), suggesting lengthening of all phases. Surprisingly, therefore, although activating E2fs contribute to progenitor expansion, robust division continues in their absence. E2f1 is essential to drive abnormal division of *Rb1*-deficient differentiating retinal neurons[14], highlighting the distinct role for activating E2fs in normal versus cancer cell cycles in the same tissue.

Mature neurons never divide, but retinal damage triggers reactive gliosis and Müller glia mitosis. We damaged retinas either genetically using homozygous *rd1* (also called $Pde6b^{-/-}$), causing photoreceptor degeneration, or by means of intravitreal toxin injection. GFAP induction and $p27^{Kip1}$ and cyclin D3 downregulation, hallmarks of reactive gliosis[15], were observed independent of *E2f1–3* (Supplementary Fig. 3a, b and data not shown). Furthermore, the proportion of dividing $BrdU^+/CRALBP^+$ Müller nuclei in the inner nuclear layer was identical in control or triple knockout toxin-treated retina (Supplementary Fig. 3c, d). Thus, both progenitors and mature glia proliferate without activating E2fs.

Some E2f targets (for example, cyclins/*Mcm3*/*Tk1*) were downregulated in $E2f1^{-/-}E2f2^{-/-}E2f3^{-/-}$ triple knockout progenitors (Fig. 2a), and we wondered if reduction of repressive E2fs (E2f4–8) might explain continued division. However, whereas *E2f7* and *E2f8* mRNA levels were slightly down in the triple knockout retina, *E2f4–6* levels were unaffected at E14, *E2f5* and *E2f6* were induced at P0 (Fig. 2a). Triple knockout fibroblasts arrest as a result of p53 induction of *Cdkn1a*, which encodes the cyclin-dependent kinase (Cdk) inhibitor $p21^{Cip1}$ (refs 4, 5). Notably, expression of *Cdkn1a* and the related *Cdkn1c* ($p57^{Kip2}$) were reduced in triple knockout retina (Fig. 2b). Myc (previously c-Myc) represses the *Cdkn1a* promoter *in vitro*[16] and its overexpression bypasses arrest caused by inhibiting E2f[17,18]. Because E2f inhibition in these studies targeted all E2fs, it is unclear if Myc overrides loss of activating E2fs alone. Whether physiological Myc levels can compensate for E2f1–3 *in vivo* is also unknown. Notably, *Mycn* and *Mycl1* levels were higher than *Myc* in E14 retina, but the reverse applied in fibroblasts (Fig. 2c). *E2f1–3* loss reduced *Myc* levels in fibroblasts, but had no effect on Myc family expression in retinal progenitors (Fig. 2c). To test Mycn function, the floxed (*f*) locus was introduced into the triple knockout background and recombination confirmed (Supplementary Fig. 4). Deleting

[1]Toronto Western Research Institute, University Health Network, Departments of Ophthalmology and Visual Science, and Laboratory Medicine and Pathobiology, University of Toronto, Ontario M5T 2S8, Canada. [2]Human Cancer Genetics Program, Department of Molecular Virology, Immunology and Medical Genetics, and Department of Molecular Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210, USA. [3]Departments of Cell Biology and Human Anatomy University of California Davis School of Medicine, Davis, California 95616, USA.

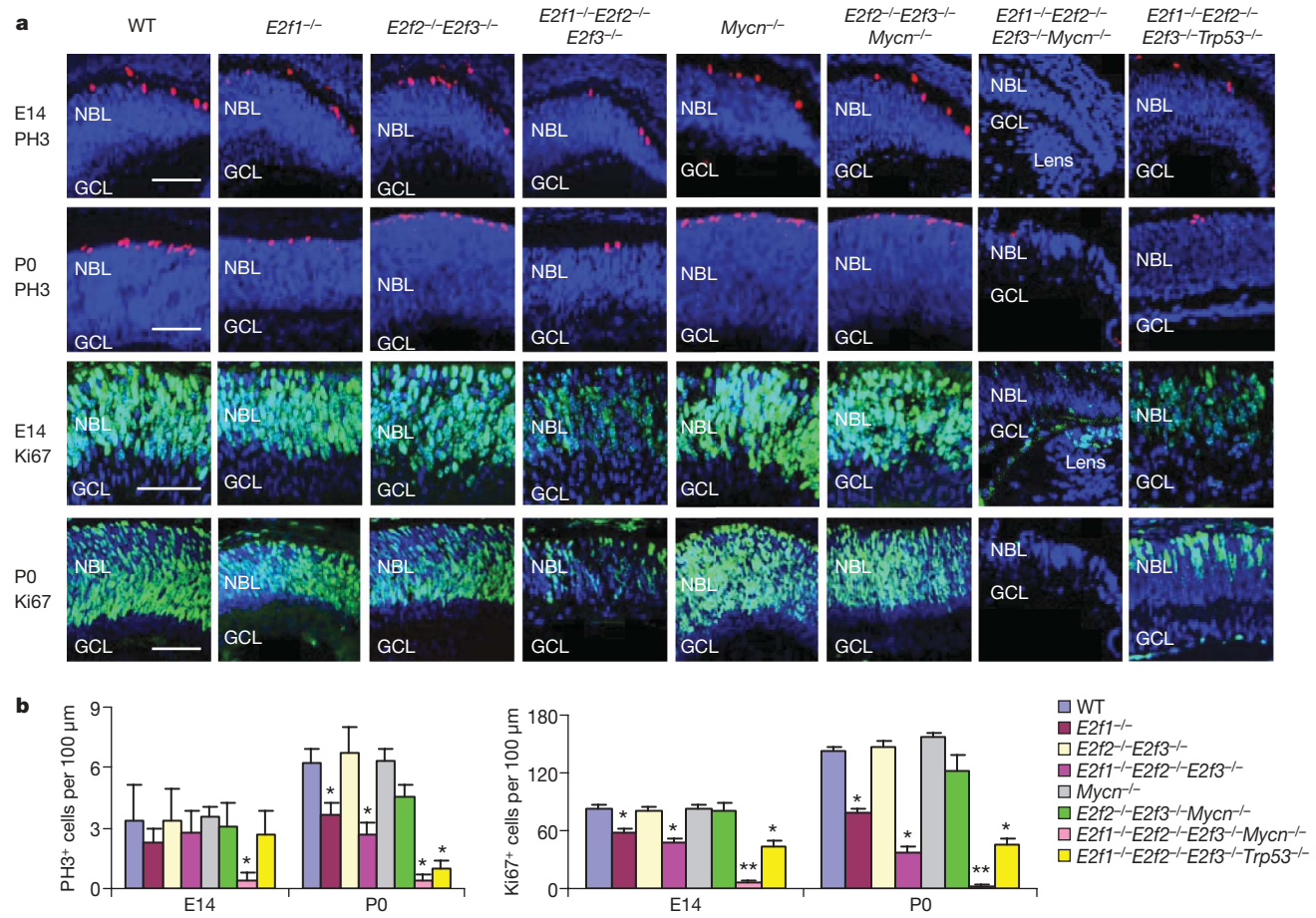



**Figure 1 | Mycn allows division without activating E2fs. a**, Retinal sections of indicated genotypes and ages were stained for nuclei (DAPI (4,6-diamidino-2-phenylindole), blue), mitoses (top two panels, PH3, red) or all dividing cells (bottom two panels, Ki67, green). Owing to the small size of the E14 *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ retina, the lens is also visible in these panels. Scale bar, 50 µm. WT, wild type. **b**, Quantification of PH3$^{+}$ and Ki67$^{+}$ cells. GCL, ganglion cell layer; NBL, neuroblastic layer. Data in **b** are mean ± s.d. and asterisks indicate significant difference from wild type ($n = 3$). Asterisk, $P < 0.05$; double asterisk, $P < 0.01$; Student's *t*-test.

*Mycn* alone, or even with *E2f2/3*, did not affect proliferation or survival (Figs 1a, b and 3a–c and data not shown; see also Supplementary Discussion on eye size). In contrast, progenitor division was markedly reduced in the *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ quadruple knockout peripheral retina with 12-fold or 67-fold fewer Ki67$^{+}$ cells than wild-type retina at E14 and P0, respectively, and the neuroblastic layer was very thin (Fig. 1a). PH3 and BrdU analyses confirmed this dramatic effect (Fig. 1a, b, and data not shown). To address cell autonomy in Mycn/E2f1–3 redundancy, Cre retrovirus co-expressing green fluorescent protein (GFP) (Supplementary Fig. 5) was injected subretinally into newborn mice and clone size analysed at P21. The average size of wild-type, *E2f2*$^{-/-}$, *E2f2*$^{-/-}$*E2f3*$^{-/-}$ and *Mycn*$^{-/-}$ clones was similar (~2.1 cells per clone); *E2f1*$^{-/-}$*E2f2*$^{-/-}$ and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ clones were smaller (~1.7 cells per clone); but *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ clones were almost always single cells (1.07 cells per clone) (Supplementary Fig. 6). Analysis of E2f target gene expression revealed that a subset of E2f targets that were reduced by *E2f1–3* loss at E14 was further reduced in the *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ retina, and a larger subset unaffected by *E2f1–3* loss was downregulated in the E14 *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ retina (Supplementary Fig. 7 and Supplementary Discussion). Notably, whereas *Cdkn1a* and *Cdkn1c*

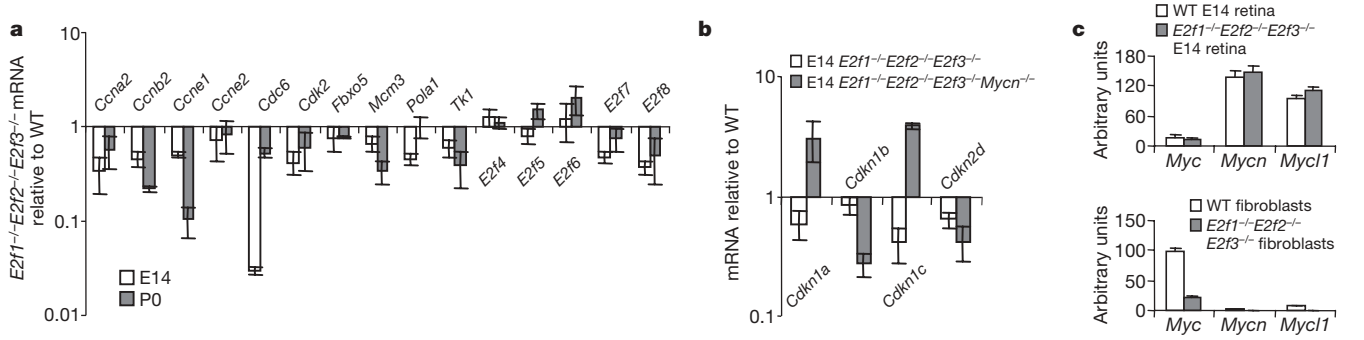



**Figure 2 | Mycn blocks Cdk inhibitor induction in E2f1–3 null progenitors. a**, Quantitative polymerase chain reaction with reverse transcription (qRT–PCR) for the indicated genes in E14 and P0 wild-type and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retinas. Triple knockout expression levels are shown relative to wild type. **b**, qRT–PCR for indicated genes in E14 wild-type, *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ retinas. Levels are plotted relative to wild type. **c**, qRT–PCR for Myc relatives in indicated cells. *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ cells were GFP-sorted from α-*CreE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ retinas or *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ fibroblasts transduced with MXIE-Cre virus. All data are mean ± s.d. ($n = 3$).

were reduced in the *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retina, they were upregulated in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ retina (Fig. 2b, Supplementary Fig. 7 and Supplementary Discussion). Thus, Mycn promotes division of *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ progenitors cell autonomously and maintains E2f target expression while repressing Cdk inhibitors.

Because *E2f1–3* are non-essential for division, we considered other roles. Activating E2fs are pro-apoptotic[19], but at E14 and E17 *E2f1–3* loss did not affect low background apoptosis, yet remarkably, apoptosis was elevated in postnatal *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retina, and could be suppressed by any single activating E2f (Fig. 3a–c and Supplementary Fig. 8). By P8, when progenitor division has ended, apoptosis was normal (Supplementary Fig. 8a–c). Labelling with TdT-mediated dUTP nick end labelling (TUNEL) and BrdU showed that apoptotic *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ cells were progenitors (Supplementary Fig. 9). This phenotype may explain the greater reduction in progenitors at P0 versus E14 (Fig. 1a, b). Thus, activating E2fs have an unexpected pro-survival role during retinal development.

In flies, E2f1 protects some irradiated cells by repressing pro-apoptotic *hid*[9], but in the *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retina the mammalian equivalents *Smac/Diablo* and *Omi/Htra2* were downregulated (Fig. 3d). The E2f target *p73* was also downregulated, but immunoreactivity for its relative p53 was elevated, and its pro-apoptotic target *Noxa* (also called Pmaip1) was induced (Fig. 3d–g). We introduced the floxed p53 locus (*Trp53*$^{f/f}$) into the α-*CreE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ background, confirmed recombination (Supplementary Fig. 10) and found that whereas there was no difference in Ki67$^{+}$, PH3$^{+}$ or BrdU$^{+}$ cells in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ versus *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Trp53*$^{-/-}$ progenitors at E14 or P0 (Fig. 1a, b, and data not shown), apoptosis was suppressed in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Trp53*$^{-/-}$ P0 progenitors (Fig. 3a–c). Although *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$*Mycn*$^{-/-}$ P0 retina had fewer total apoptotic cells (Fig. 3a–c), this was due to fewer retinal progenitors; thus, the proportion of apoptotic (TUNEL$^{+}$) dividing (Ki67$^{+}$) progenitors at P0 was not reduced by *Mycn* loss, but was slightly increased (Supplementary Fig. 11), consistent with the slightly elevated Noxa expression relative to the *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retina (Supplementary Fig. 7). Thus, unlike in fibroblasts[4], p53 does not reduce division in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ progenitors, but drives apoptosis, providing and explaining the first pro-survival role for activating E2fs in normal cells.

How do E2fs constrain p53 pro-apoptotic activity? Aspp proteins bind p53 and promote its pro-apoptotic function, but they are

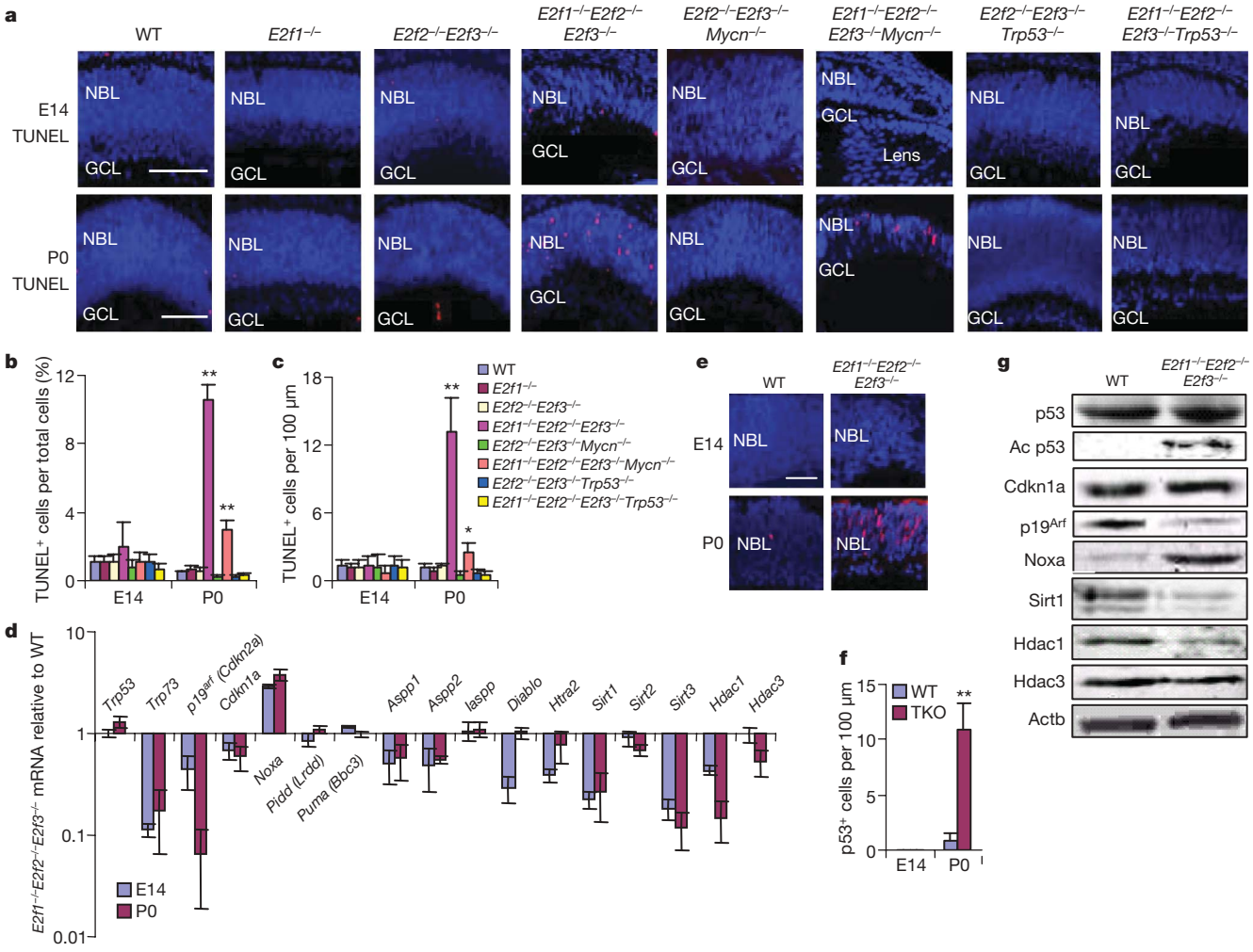



**Figure 3 | A pro-survival role for activating E2fs.** **a**, Retinal sections of indicated genotypes and ages were stained for nuclei (DAPI, blue) and apoptosis (TUNEL, red). **b**, **c**, Quantification of TUNEL$^{+}$ cells: **b**, proportion; **c**, number per 100 μm. **d**, qRT–PCR analysis of indicated genes in E14 and P0 wild-type and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retinas. Expression levels in the triple knockout are shown relative to those in wild-type cells. **e**, Retinal sections of the indicated genotypes and ages were stained for nuclei (DAPI, blue) and p53 (red). **f**, Quantification of p53$^{+}$ cells. **g**, Cell lysates of P0 wild-type and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retinas were probed with the indicated antibodies on western blots. *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ cells were dissected from α-*CreE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ peripheral retina. Scale bars in **a**, **e** are 50 μm. GCL, ganglion cell layer; NBL, neuroblastic layer. Data in **b**–**d**, **f** are mean ± s.d. Asterisks in **b**, **c** indicate significant difference from wild type ($n = 3$). Asterisk, $P < 0.05$; double asterisk, $P < 0.01$; Student's *t*-test.

E2f-induced[20,21] and were downregulated in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ progenitors; the Aspp inhibitor, *Iaspp* (also called Ppp1r13l), was unaffected (Fig. 3d). p19$^{Arf}$ stabilizes p53 by inhibiting Mdm2, and *E2f3* loss de-represses *p19*$^{Arf}$ transcription in fibroblasts[22], but we observed less *p19*$^{Arf}$ mRNA and protein in triple knockout progenitors (Fig. 3d, g). *E2f1–3* loss did not affect *Trp53* mRNA, and total p53 protein levels were increased only ~1.5-fold in *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ retina (Fig. 3d, g), thus the 13-fold increase in immunoreactivity may reflect epitope exposure (Fig. 3e, f), perhaps through posttranslational modification[23]. *E2f1–3* loss did not induce p53 phosphorylation (data not shown) but increased acetylation considerably (Fig. 3g). Acetylation is essential for p53 activity[24], is mediated by Tip60, Pcaf and Cbp/p300 and is reversed by Hdacs and Sirt1[23]. Notably, *Sirt1*, *Sirt2*, *Sirt3*, *Hdac1* and *Hdac3* mRNA was reduced in triple knockout retina, as were protein levels of Sirt1 and Hdac1 (Fig. 3d, g). Hdac1 and Hdac3 downregulation was greater at P0 than at E14, correlating with p53 immunodetection at P0 (Fig. 3d–f). *Sirt1* is a direct E2f1 target[25], but whether other activating E2fs sustain its expression is unknown, and a functional E2f–Sirt1–p53 axis has not been established. Electroporating Cre plasmid into *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ P0 retinas induced apoptosis, but strikingly concomitant Sirt1 expression blocked cell death (Fig. 4a, b). Moreover, treatment of pregnant dams from E16 with the Sirt1 agonist resveratrol blocked >70% of *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{-/-}$ progenitor apoptosis at P0, and this correlated with p53 deacetylation and blockade of p53-dependent Noxa induction, without affecting p19$^{Arf}$ (Supplementary Fig. 12). Just as p53 loss did not influence division (Fig. 1a, b), resveratrol did not alter cyclin levels or proliferation (Supplementary Figs 12d and 13). In summary, activating E2fs interchangeably induce Sirt1 to block p53 acetylation and apoptosis. High levels of E2f have long been known to activate p53 (ref. 8) and our findings now connect low E2f to p53 activation, but through an entirely distinct mechanism (Fig. 4c).

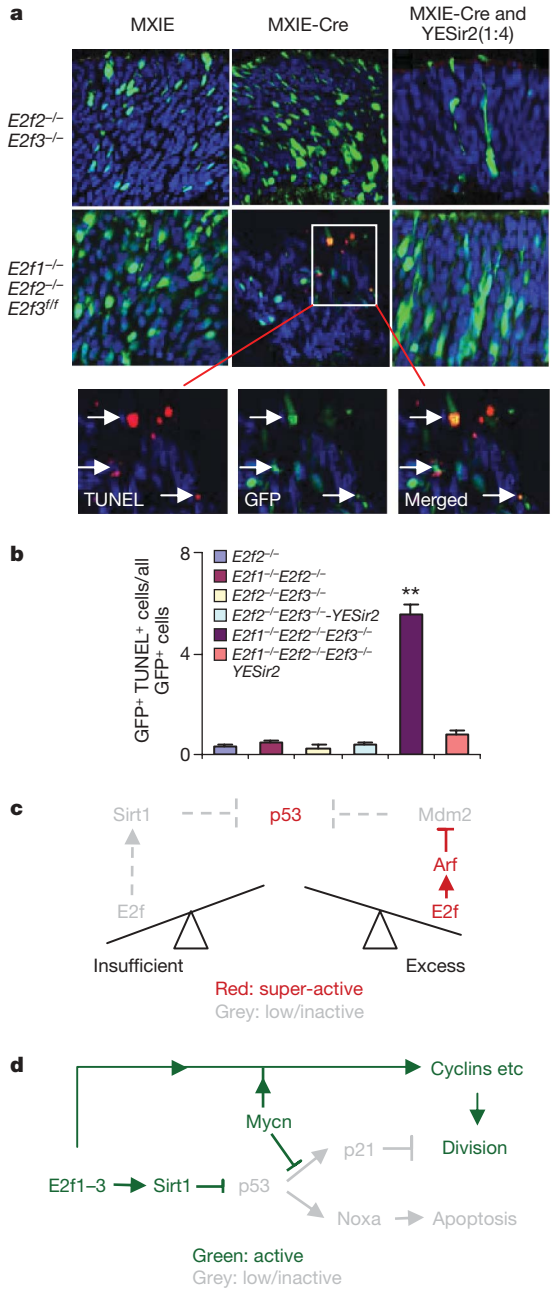



**Figure 4 | E2fs promote survival through Sirt1-mediated p53 deacetylation. a**, MXIE plasmid, MXIE-Cre plasmid, or MXIE-Cre plasmid plus YESir2 plasmid expressing Sirt1 (ref. 29) were injected subretinally and electroporated into the P0 retinas of the indicated genotypes and analysed at P2. GFP (green) marks transfected cells and sections were also stained for nuclei (DAPI, blue) and apoptosis (TUNEL, red). Arrows show examples of GFP$^+$TUNEL$^+$ cells. **b**, Quantification of GFP$^+$TUNEL$^+$ cells as a fraction of all GFP$^+$ cells in clones of the indicated genotype. Data are mean ± s.d. ($n = 3$). Double asterisk, $P < 0.01$ relative to wild type; Student's *t*-test. **c**, Insufficient or excess E2f activates p53, but by distinct mechanisms. Red, high/super-active; grey, low/inactive. **d**, An integrated E2f–Myc–Sirt1 network promotes division and survival in retinal progenitors. Green and grey indicate active or inactive factors/functions, respectively.

Activating E2fs are critical for division in flies and mammalian fibroblasts, where they counteract repressive E2f or p53, respectively[3–7]. We provide the first *in vivo* evidence that E2f1–3 are not indispensable in every context. Mycn drives E2f-independent division, maintaining expression of E2f targets and preventing p53-mediated *Cdkn1a* induction. *In vitro* overexpression studies suggested that Myc overcomes E2f repression[17,18] but did not reveal whether normal Myc levels substitute for activating E2fs *in vivo*. Our genetic strategies reveal that physiological levels compensate for the activating E2fs cell autonomously. Other studies highlight redundancy among related cell cycle regulators[26]. Our results indicate redundancy between unrelated cell cycle regulators.

We also describe an unexpected pro-survival activity for E2f1–3 *in vivo*. Loss of any two was harmless, but removing all three impaired progenitor survival. Thus, activating E2fs have interchangeable pro-survival roles—not only E2f1 and not only in irradiated cells[9,10,12]. We define the first direct mechanism linking activating E2fs to mammalian cell survival through Sirt1 and p53. E2f1–3 redundantly regulates other deacetylases (Fig. 3d), which may also inhibit p53. Integrating the Sirt1–p53 and Mycn data exposes a network that coordinates progenitor proliferation and survival (Fig. 4d). The novel E2f–Sirt–p53 axis could also promote survival in p53$^+$ tumour cells. Irrespective, its importance for survival in normal cells underscores the need for care in applying E2f inhibitors to treat human disease.

## METHODS SUMMARY

For reactive gliosis 1.0 µl of 2 mM domoic acid and 7 mM ouabain plus 1 mM BrdU was injected into the eyes of P16 ketamine/xylazine anaesthetized mice[15]. For resveratrol, timed pregnant female mice were injected daily with 4 mg kg$^{-1}$ of body weight (catalogue number 60512A, AKSci) or 0.1% DMSO intraperitoneally from E16. Virus injection and electroporation were as described[27]. Fixation, antibody labelling, TUNEL and quantification of sections, or dissociation and quantification of cells, and RNA and protein analysis were essentially as described[14,28]. The PALM microlaser system was used for laser capture microdissection (LCM) following the manufacturer's recommendations. For flow cytometry *E2f3*$^{f/f}$ and *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{f/f}$ fibroblasts were infected with MXIE or MXIRE-Cre retrovirus. E14 retinas were dissected and dissociated. GFP$^+$ cells were sorted using a FACSAria (Becton Dickinson). For cell cycle analysis, cells were fixed in ice-cold 80% ethanol, counterstained with propidium iodide and analysed with a BD FACSCalibur system. Data were collected using CELLFIT software.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 13 July; accepted 25 September 2009.**


1. Rowland, B. D. & Bernards, R. Re-evaluating cell-cycle regulation by E2Fs. *Cell* **127,** 871–874 (2006).
2. Pacal, M. & Bremner, R. Insights from animal models on the origins and progression of retinoblastoma. *Curr. Mol. Med.* **6,** 759–781 (2006).
3. van den Heuvel, S. & Dyson, N. J. Conserved functions of the pRB and E2F families. *Nature Rev. Mol. Cell Biol.* **9,** 713–724 (2008).
4. Wu, L. *et al.* The E2F1–3 transcription factors are essential for cellular proliferation. *Nature* **414,** 457–462 (2001).
5. Timmers, C. *et al.* E2f1, E2f2, and E2f3 control E2F target expression and cellular proliferation via a p53-dependent negative feedback loop. *Mol. Cell. Biol.* **27,** 65–78 (2007).
6. Sharma, N. *et al.* Control of the p53-p21CIP1 axis by E2f1, E2f2, and E2f3 is essential for G1/S progression and cellular transformation. *J. Biol. Chem.* **281,** 36124–36131 (2006).
7. Frolov, M. V. *et al.* Functional antagonism between E2F family members. *Genes Dev.* **15,** 2146–2160 (2001).
8. Iaquinta, P. J. & Lees, J. A. Life and death decisions by the E2F transcription factors. *Curr. Opin. Cell Biol.* **19,** 649–657 (2007).
9. Moon, N. S. *et al. Drosophila* E2F1 has context-specific pro- and antiapoptotic properties during development. *Dev. Cell* **9,** 463–475 (2005).
10. Wikonkal, N. M. *et al.* Inactivating E2f1 reverts apoptosis resistance and cancer sensitivity in Trp53-deficient mice. *Nature Cell Biol.* **5,** 655–660 (2003).
11. Xie, W. *et al.* Novel link between E2F1 and Smac/DIABLO: proapoptotic Smac/DIABLO is transcriptionally upregulated by E2F1. *Nucleic Acids Res.* **34,** 2046–2055 (2006).
12. Berton, T. R., Mitchell, D. L., Guo, R. & Johnson, D. G. Regulation of epidermal apoptosis and DNA repair by E2F1 in response to ultraviolet B radiation. *Oncogene* **24,** 2449–2460 (2005).
13. Livesey, F. J. & Cepko, C. L. Vertebrate neural cell-fate determination: lessons from the retina. *Nature Rev. Neurosci.* **2,** 109–118 (2001).
14. Chen, D. *et al.* Rb-mediated neuronal differentiation through cell-cycle-independent regulation of E2f3a. *PLoS Biol.* **5,** e179 (2007).
15. Dyer, M. A. & Cepko, C. L. Control of Muller glial cell proliferation and activation following retinal injury. *Nature Neurosci.* **3,** 873–880 (2000).
16. Claassen, G. F. & Hann, S. R. A role for transcriptional repression of p21CIP1 by c-Myc in overcoming transforming growth factor β-induced cell-cycle arrest. *Proc. Natl Acad. Sci. USA* **97,** 9498–9503 (2000).
17. Alevizopoulos, K., Vlach, J., Hennecke, S. & Amati, B. Cyclin E and c-Myc promote cell proliferation in the presence of p16INK4a and of hypophosphorylated retinoblastoma family proteins. *EMBO J.* **16,** 5322–5333 (1997).
18. Santoni-Rugiu, E., Falck, J., Mailand, N., Bartek, J. & Lukas, J. Involvement of Myc activity in a $G_1$/S-promoting mechanism parallel to the pRb/E2F pathway. *Mol. Cell. Biol.* **20,** 3497–3509 (2000).
19. DeGregori, J. & Johnson, D. G. Distinct and overlapping roles for E2F family members in transcription, proliferation and apoptosis. *Curr. Mol. Med.* **6,** 739–748 (2006).
20. Chen, D., Padiernos, E., Ding, F., Lossos, I. S. & Lopez, C. D. Apoptosis-stimulating protein of p53–2 (ASPP2/53BP2L) is an E2F target gene. *Cell Death Differ.* **12,** 358–368 (2005).
21. Fogal, V. *et al.* ASPP1 and ASPP2 are new transcriptional targets of E2F. *Cell Death Differ.* **12,** 369–376 (2005).
22. Aslanian, A., Iaquinta, P. J., Verona, R. & Lees, J. A. Repression of the Arf tumor suppressor by E2F3 is required for normal cell cycle kinetics. *Genes Dev.* **18,** 1413–1422 (2004).
23. Bode, A. M. & Dong, Z. Post-translational modification of p53 in tumorigenesis. *Nature Rev. Cancer* **4,** 793–805 (2004).
24. Tang, Y., Zhao, W., Chen, Y., Zhao, Y. & Gu, W. Acetylation is indispensable for p53 activation. *Cell* **133,** 612–626 (2008).
25. Wang, C. *et al.* Interactions between E2F1 and SirT1 regulate apoptotic response to DNA damage. *Nature Cell Biol.* **8,** 1025–1031 (2006).
26. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nature Rev. Cancer* **9,** 153–166 (2009).
27. Livne-bar, I. *et al.* Chx10 is required to block photoreceptor differentiation but is dispensable for progenitor proliferation in the postnatal retina. *Proc. Natl Acad. Sci. USA* **103,** 4988–4993 (2006).
28. Chen, D. *et al.* Cell-specific effects of RB or RB/p107 loss on retinal development implicate an intrinsically death-resistant cell-of-origin in retinoblastoma. *Cancer Cell* **5,** 539–551 (2004).
29. Vaziri, H. *et al.* hSIR2(SIRT1) functions as an NAD-dependent p53 deacetylase. *Cell* **107,** 149–159 (2001).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank L. van Parijs and R. Weinberg for plasmids; R. Eisenman for mice; L. Penn for advice on Myc; and M. Cayouette and J. Wrana for comments. This work was supported by a grant from the Canadian Institutes for Health Research to R.B. (MOP-74570).

**Author Contributions** D.C. and R.B. designed the study and interpreted data. D.C. performed all the experiments and was aided in viral and electroporation assays by M.P. P.W., G.L. and P.S.K. provided reagents including mice. R.B. wrote the paper and all authors contributed to editing.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.B. (rbremner@uhnres.utoronto.ca).

## METHODS

**Mice, genotyping and resveratrol injection.** Mice were treated according to institutional and national guidelines. α-*Cre* (P. Gruss)[30], ROSA26R (Jackson Laboratories)[31], $E2f1^{-/-}$ mice (M. Greenberg) and $E2f2^{-/-}$, $E2f3^{f/f}$ (ref. 32), $Trp53^{f/f}$ (ref. 33), $Mycn^{f/f}$ (ref. 34) and $Pde6b^{-/-}$ (referred to as *rd1*, Jackson Laboratories) mice were maintained on a mixed (NMRI × C57/Bl × FVB/N × 129sv) background. Genotyping was performed according to published protocols. For resveratrol assays, timed pregnant female mice were injected daily with 4 mg $kg^{-1}$ of body weight resveratrol (catalogue number 60512A, AKSci) or 0.1% DMSO intraperitoneally from E16.

**Immunofluorescence.** Eyes were fixed in 4% paraformaldehyde for 1 h at 4 °C, embedded in OCT (TissueTek 4583), frozen on dry ice and cut into 12 μm sections on Superfrost slides. BrdU (100 μg $g^{-1}$ body weight) was injected intraperitoneally into mice 2 h before they were killed. $BrdU^+$ cells were detected using a biotin-conjugated sheep polyclonal antibody (1:500, Maine Biotechnology Services). Other antibodies were CRALBP (J. Saari), cyclin D3 (Santa Cruz, SC-6283), E2f3 (Upstate, 05-551), GFP (Molecular Probe, A11122), GFAP (Sigma, G9269), Ki67 (BD Biosciences Pharmingen, 550609), Ki67 (Neomarkers, RM-9106S), $p27^{Kip1}$ (Santa Cruz, SC-528), Trp53 (NCL-p53-CM5p; Novocastra) and phospho-histone H3 (Upstate Biotechnology, 06-570). TUNEL and antigen retrieval was performed as described[14]. For double labelling of BrdU and other markers, sections were first stained for the other marker, and fixed again with 70% ethanol/10% acetic acid for 5 min. Sections were washed, treated with 2 N HCl, and stained for BrdU. Primary antibodies were visualized using goat anti-mouse Alexa-488 or Alexa-568, goat anti-rabbit Alexa-488 or Alexa-568, donkey anti-goat Alexa-568, Streptavidin Alexa-488 or Alexa-568 (1:1,000; Molecular Probes) and Streptavidin-HRP-DAB. Nuclei were counterstained with 4,6-diamidino-2-phenylindole (DAPI; Sigma). Labelled cells were visualized using a Zeiss Axioplan-2 microscope with Plan Neofluar objectives and images captured with a Zeiss AxionCam camera. For double-labelled samples, confocal images were obtained with a Zeiss LSM 5.0 Laser Scanning microscope. For quantification of stained sections, the retina was separated into bins[28]. Measurements were performed with Axiovison software. Quantification used horizontal sections containing the optic nerve. At least three sections per eye and three eyes from different litters were counted. Statistical analysis used the Student's *t*-test. *P* values are based on two-sided hypothesis testing.

**Reactive gliosis.** 1.0 μl of 2 mM domoic acid and 7 mM ouabain plus 1 mM BrdU was injected into the eyes of ketamine/xylazine anaesthetized P16 mice[15].

**Laser capture microdissection.** Retinal cells were dissected from 12 μm sections using the PALM microlaser system, catapulted into 10 μl buffer (1 mM EDTA; 20 mM Tris (pH 8)) containing 2 mg $ml^{-1}$ proteinase K, incubated at 55 °C overnight, and proteinase K inactivated at 99 °C for 10 min. 5 μl was used for PCR.

**Flow cytometry.** $E2f3^{f/f}$ and $E2f1^{-/-}E2f2^{-/-}E2f3^{f/f}$ mouse fibroblasts were infected with MXIE or MXIRE-Cre retrovirus. E14 retinas were dissected and dissociated. $GFP^+$ cells were sorted using a FACSAria (Becton Dickinson). For cell cycle analysis, cells were fixed in ice-cold 80% ethanol, counterstained with propidium iodide and analysed with a BD FACSCalibur system. Data were collected using CELLFIT software.

**RT–PCR and western blotting.** Total RNA from fibroblast cells, peripheral retina or GFP-sorted cells was prepared using the RNeasy Mini kit (Qiagen), and treated with DNA-free (Ambion) according to the manufacturer's instructions. First-strand cDNA was synthesized from 0.2–0.5 μg using SuperScript II (Invitrogen); primers are listed in Supplementary Table 1. An Applied Biosystems PRISM 7900HT and SYBR Green PCR master mix was used for real-time PCR. Tests were run in duplicate on three biological samples. Values were normalized to β-actin.

For western blots, peripheral mouse retinas were homogenized with a 30 gauge needle 5–10 times in 1× cell lysis buffer (Cell Signaling) with 0.1 mM PMSF, 1 μg $ml^{-1}$ aprotinin, 1 μg $ml^{-1}$ leupeptin, 2 mM $NaN_3$, 10 μM trichostatin A and 5 mM niacinamide. Proteins were separated by 12% SDS–PAGE, transferred to nitrocellulose and analysed by Li-Cor system (Lincoln) with antibodies against p53 (NCL-p53-CM5p; Novocastra), acetylated p53 (ab61241; Abcam), phosphorylated p53 (9286S; Cell Signaling), Sirt1 (07-131; Upstate), $p21^{Cip1}$ (M-19; Santa Cruz), $p19^{Arf}$ (NB200-106; Novus Biologicals), Noxa (IMG451; Imgenex), HDAC1 (H-51; Santa Cruz), HDAC3 (H-99; Santa Cruz) and β-actin (A5441, Sigma).

**Retroviral injection and *in vivo* electroporation.** Cre recombinase cDNA (L. van Parijs) was cloned into the BglII site of MXIE and retrovirus generated in Phoenix-eco cells as described[27]. Recombinase activity was confirmed in ROSA26R mice (Supplementary Fig. 5) and by PCR of LCM-dissected single cells (M.P. and R.B., unpublished data). P0 pups were anaesthetized on ice. 2 μl retrovirus was injected into the subretinal space of right (MXIE) or left eye (MXIE-Cre) through a small limbal incision. P21 retinas were stained for GFP. Note that if an $E2f1^{-/-}E2f2^{-/-}E2f3^{-/-}$ progenitor dies after introduction of Cre at P0, then by definition it will not generate any post-mitotic cells and thus will not generate a clone that can be counted at P21. Only the $E2f1^{-/-}E2f2^{-/-}E2f3^{-/-}$ progenitors that survive contribute to clones that are counted at P21. It is impossible to estimate accurately using this assay the fraction of progenitors that die because the number of clones will vary depending on technical variance associated with injection technique. This assay measures the number of cells per clone rather than the number of clones. That is, it measures the ability of surviving progenitors to divide (clone size), and not the degree of progenitor survival.

For electroporation, 1 μl of DNA (1–4 μg $\mu l^{-1}$) was injected into the subretinal space of P0 pups using a Hamilton syringe, and square electric pulses (80 V; five 50-ms with 950-ms intervals) applied with tweezer-type electrodes (CUY21 EDIT Electroporator).

**Selection of E2f targets.** E2f targets for qRT–PCR analysis were selected from primary papers and reviews, focusing mainly on chromatin immunoprecipitation studies[11,21,22,25,35–41]. Their interaction with and/or regulation by Myc family proteins (summarized in Supplementary Fig. 7) was then based on surveying both the literature[42–45] and a database of >1,600 Myc targets (http://www.myc-cancer-gene.org/)[46].


30. Marquardt, T. *et al.* Pax6 is required for the multipotent state of retinal progenitor cells. *Cell* **105,** 43–55 (2001).
31. Soriano, P. Generalized *lacZ* expression with the ROSA26 Cre reporter strain. *Nature Genet.* **21,** 70–71 (1999).
32. Leone, G. *et al.* Myc requires distinct E2F activities to induce S phase and apoptosis. *Mol. Cell* **8,** 105–113 (2001).
33. Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nature Genet.* **29,** 418–425 (2001).
34. Knoepfler, P. S., Cheng, P. F. & Eisenman, R. N. N-myc is essential during neurogenesis for the rapid expansion of progenitor cell populations and the inhibition of neuronal differentiation. *Genes Dev.* **16,** 2699–2712 (2002).
35. Bracken, A. P., Ciro, M., Cocito, A. & Helin, K. E2F target genes: unraveling the biology. *Trends Biochem. Sci.* **29,** 409–417 (2004).
36. Xu, X. *et al.* A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.* **17,** 1550–1561 (2007).
37. Ozono, E. *et al.* E2F-like elements in p27(Kip1) promoter specifically sense deregulated E2F activity. *Genes Cells* **14,** 89–99 (2009).
38. Weinmann, A. S., Bartley, S. M., Zhang, T., Zhang, M. Q. & Farnham, P. J. Use of chromatin immunoprecipitation to clone novel E2F target promoters. *Mol. Cell. Biol.* **21,** 6820–6832 (2001).
39. Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H. & Farnham, P. J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16,** 235–244 (2002).
40. Christensen, J. *et al.* Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. *Nucleic Acids Res.* **33,** 5458–5470 (2005).
41. Lyons, T. E., Salih, M. & Tuana, B. S. Activating E2Fs mediate transcriptional regulation of human E2F6 repressor. *Am. J. Physiol. Cell Physiol.* **290,** C189–C199 (2006).
42. Mao, D. Y. *et al.* Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. *Curr. Biol.* **13,** 882–886 (2003).
43. Bieda, M., Xu, X., Singer, M. A., Green, R. & Farnham, P. J. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res.* **16,** 595–605 (2006).
44. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–1061 (2008).
45. Zeller, K. I. *et al.* Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl Acad. Sci. USA* **103,** 17834–17839 (2006).
46. Zeller, K. I., Jegga, A. G., Aronow, B. J., O'Donnell, K. A. & Dang, C. V. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. *Genome Biol.* **4,** R69 (2003).

LETTERS

# E2f1–3 switch from activators in progenitor cells to repressors in differentiating cells

Jean-Leon Chong[1,2,3]*, Pamela L. Wenzel[1,2,3]*†, M. Teresa Sáenz-Robles[4]*, Vivek Nair[1,2,3], Antoney Ferrey[1,2,3], John P. Hagan[1,3], Yorman M. Gomez[1,2,3], Nidhi Sharma[1,2,3], Hui-Zi Chen[1,2,3], Madhu Ouseph[1,2,3], Shu-Huei Wang[1,2,3], Prashant Trikha[1,2,3], Brian Culp[1,2,3], Louise Mezache[1,2,3], Douglas J. Winton[5], Owen J. Sansom[6], Danian Chen[7], Rod Bremner[7], Paul G. Cantalupo[4], Michael L. Robinson[8], James M. Pipas[4] & Gustavo Leone[1,2,3]

**In the established model of mammalian cell cycle control, the retinoblastoma protein (Rb) functions to restrict cells from entering S phase by binding and sequestering E2f activators (*E2f1*, *E2f2* and *E2f3*), which are invariably portrayed as the ultimate effectors of a transcriptional program that commit cells to enter and progress through S phase[1,2]. Using a panel of tissue-specific *cre*-transgenic mice and conditional *E2f* alleles we examined the effects of *E2f1*, *E2f2* and *E2f3* triple deficiency in murine embryonic stem cells, embryos and small intestines. We show that in normal dividing progenitor cells E2f1–3 function as transcriptional activators, but contrary to the current view, are dispensable for cell division and instead are necessary for cell survival. In differentiating cells E2f1–3 function in a complex with Rb as repressors to silence E2f targets and facilitate exit from the cell cycle. The inactivation of *Rb* in differentiating cells resulted in a switch of E2f1–3 from repressors to activators, leading to the superactivation of E2f responsive targets and ectopic cell divisions. Loss of *E2f1–3* completely suppressed these phenotypes caused by *Rb* deficiency. This work contextualizes the activator versus repressor functions of E2f1–3 *in vivo*, revealing distinct roles in dividing versus differentiating cells and in normal versus cancer-like cell cycles.**

E2fs function as transcription factors, with E2f1–3 as activators and E2f4–8 as repressors[3–8]. Although it is a maxim of mammalian cell cycle regulation that the E2f1–3 activator subclass is required for cell proliferation, the evidence for this is based almost exclusively on *in vitro* studies using cells derived from murine and human tissues or on the *in vivo* analysis of *Rb* mutant mice[1,2]. Other experiments, however, suggest that these E2fs can also function as repressors in complex with Rb[9–11], yet the relative contribution of activation versus repression and the physiological contexts in which these contrary E2f functions are used remain unclear.

To explore the functions of the E2f activator subclass, we derived $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/-}$ embryonic stem (ES) cells (Supplementary Fig. 1a, b) and compared the consequences of inactivating the conditional $E2f3^{loxP}$ allele in ES cells and $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ mouse embryonic fibroblasts (MEFs). The expression of *E2f1*, *E2f2* and *E2f3* in wild-type ES cells was generally higher than in MEFs and the loading of E2f3 protein on classic E2f target promoters was comparable between the two proliferating cell types (Supplementary Fig. 2a–c). Consistent with previous observations, the ablation of *E2f1–3* in MEFs with standard *cre*-expressing vectors led to the induction of p53 activity, the loading of E2f4–p130 repressor complexes on E2f target promoters and a marked decrease in E2f target expression (Fig. 1a and Supplementary Fig. 3a–c)[5–7]. Consequently, triply deficient MEFs underwent a complete cell cycle arrest (Fig. 1b)[5–7]. In contrast, $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ ES cells failed to activate p53 or form E2f4–p130 repressive complexes, and as a result, E2f target expression was unaffected and cells proliferated equally well as $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/-}$ double knockout control cells (Fig. 1b and Supplementary Fig. 3a–c).

We then evaluated whether triply-deficient ES cells could proliferate *in vivo*. Subcutaneous injection of $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ ES cells into athymic nude mice yielded efficient teratoma formation, producing mesoderm, endoderm and ectoderm at a rate similar to $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/-}$ double knockout ES cell lines (Fig. 1c and Supplementary Fig. 4a, b). Moreover, from $E2f1^{+/-}E2f2^{-/-}E2f3^{+/-}$ intercrosses we recovered the expected number of live $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ triple knockout embryos as late as embryonic day (E)9.5, but none was recovered past E11.5 (Fig. 1d and data not shown). The live E9.5 $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ embryos appeared morphologically normal by gross and histological examination (Fig. 1e and data not shown). Although cell proliferation was normal in most tissues, there was evidence of decreased proliferation and increased apoptosis in the myocardium and the first branchial arch of $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ embryos (Supplementary Fig. 5a–d). These latter observations are consistent with heart defects found in *E2f3* singly-deleted adult mice[12].

To explore whether E2f1–3 might have cell-cycle-related functions in tissues that arise later in embryonic and postnatal development, we exploited the highly organized cellular architecture of the small intestine. Maintenance of structural and functional integrity of the small intestine requires continuous epithelial regeneration[13]. Intestinal stem cells are housed at the base of crypts of Lieberkühn and give rise to transit-amplifying cells. As these cells migrate up from the base and into the finger-like extensions called villi, they exit the cell cycle and differentiate[13]. Western blot assays showed that *E2f1*, *E2f2* and both isoforms of *E2f3* (E2f3a and E2f3b) are expressed in the crypt and villus (Supplementary Fig. 6). We used *Ah-cre* mice[14] to ablate *E2f1–3* in the small intestine *in utero* or in adult mice ($Ah\text{-}creE2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ triple knockout). Induction of *Ah-cre* expression by intraperitoneal injection of β-napthoflavone (β-NF) led to the efficient deletion of $E2f3^{loxP}$ in crypt stem cells and transit-amplifying cells by 1 day after injection, and in the entire

[1]Department of Molecular Virology, Immunology and Medical Genetics, College of Medicine, [2]Department of Molecular Genetics, College of Biological Sciences, [3]Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210, USA. [4]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA. [5]Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK. [6]The Beatson Institute for Cancer Research, Glasgow G61 1BD, UK. [7]Toronto Western Research Institute, University Health Network, Departments of Ophthalmology and Visual Science, and Laboratory Medicine and Pathobiology, University of Toronto, Ontario M5T 2S8, Canada. [8]Department of Zoology, Miami University, Oxford, Ohio 45056, USA. †Present address: Division of Pediatric Hematology/Oncology, Children's Hospital Boston; Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

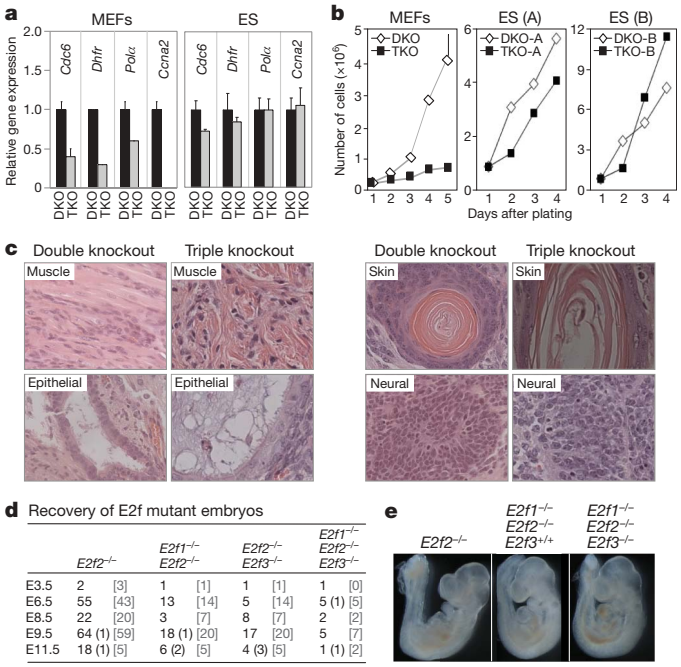



**Figure 1 | Cell proliferation in the absence of *E2f1–3*.** **a**, Expression of E2f-regulated genes was measured by real-time RT–PCR in proliferating ES and MEF cells with the indicated genotypes (primer information is provided in Supplementary Fig. 19). DKO, $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/-}$ (double knockout); TKO, $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ (triple knockout). Error bars represent standard deviation from samples analysed in triplicate. **b**, Growth curves of two sets of $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/-}$ (DKO) and $E2f1^{-/-}E2f2^{-/-}E2f3^{\Delta/-}$ (TKO) ES cell clones (A and B; right two panels), and control-vector treated $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ (DKO, left panel) and *cre*-vector treated $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ MEFs (TKO, left panel). **c**, Double knockout and triple knockout ES cells were injected underneath the skin of athymic nude mice and teratomas were harvested, sectioned and stained with haematoxylin and eosin. Representative tissues of double knockout and triple knockout teratomas include muscle (mesoderm), respiratory epithelium (endoderm), skin and neural cells (ectoderm). **d**, Embryos derived from intercrosses between $E2f1^{+/-}E2f2^{-/-}E2f3^{+/-}$ mice were collected at various time points during pregnancies. Shown are the total number of embryos collected, with the number of dead embryos (round brackets) and total number of embryos expected (square brackets) also indicated. **e**, Representative E9.5 embryos were photographed immediately upon collection.

intestinal epithelium within 3–4 days (crypt and villus; Supplementary Fig. 7a–c). Loss of *E2f1–3* did not result in a compensatory increase of other E2f family members, except for a modest increase in *E2f8* (Supplementary Fig. 7d). Whether $E2f3^{loxP}$ was deleted *in utero* at E15.5 or in the adult at 2 months of age, the architecture of $Ah\text{-}creE2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ small intestines remained relatively intact and animals were asymptomatic for 90 days after β-NF administration (Fig. 2a and Supplementary Fig. 8a, b). Cell-type-specific marker analysis demonstrated that all differentiated epithelial cell types were appropriately represented in β-NF treated $Ah\text{-}creE2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ (*E2f1–3*-deficient) small intestines (Fig. 2b and Supplementary Fig. 9). Remarkably, cell proliferation was identical in *E2f1–3*-deficient and control intestines (Fig. 2c); however, we noted a marked increase in the phosphorylation of H2AX (γ-H2AX) and ataxia telangiectasia (P-ATM$^{1981}$) proteins in *E2f1–3*-deficient crypts and villi (Fig. 2d, e and Supplementary Fig. 10a). A parallel analysis of retinal[15] and lens (P.L.W., unpublished observations) progenitors also revealed increased γ-H2AX staining in *E2f1–3*-deficient samples (Supplementary Fig. 10b, c). Together, these observations suggest that counter to current dogma, E2f1–3 are dispensable for the proliferation of ES cells and their mesodermal, endodermal and ectodermal derivatives, and for the proliferation of cells in at least some adult tissues.

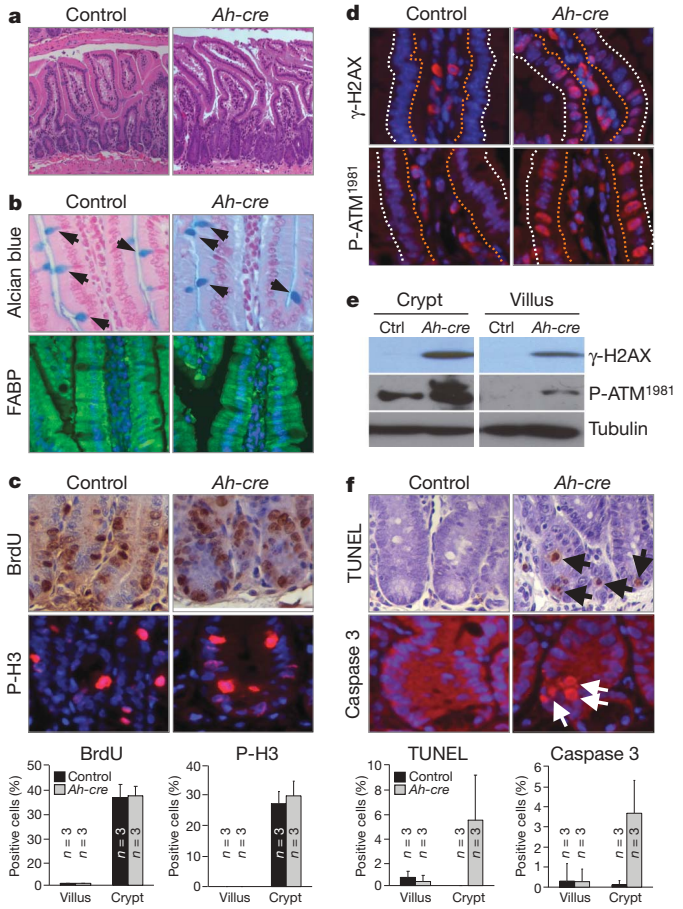



**Figure 2 | Apoptosis of crypt intestinal cells in the absence of *E2f1*, *E2f2* and *E2f3*.** **a**, Haematoxylin-and-eosin-stained sections from $E2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ (control) and $Ah\text{-}creE2f1^{-/-}E2f2^{-/-}E2f3^{loxP/loxP}$ (*Ah-cre*) intestines after 90 days of β-NF administration. **b**, Analysis of cell differentiation in control and *Ah-cre* small intestines. Goblet cells were identified by Alcian blue staining (arrows point to positive-stained goblet cells); absorptive cells were identified by anti-fatty acid binding protein (FABP, green) antibodies; 4,6-diamidino-2-phenylindole (DAPI; blue) was used for staining nuclei. **c**, BrdU (brown) and phosphorylated histone H3 (P-H3, red) immunohistochemical staining was performed on small intestine sections from β-NF-injected control and *Ah-cre* mice. Quantification of BrdU- and phosphorylated-histone-H3-positive cells in crypts and villi. $n = 3$, 3 different animals with the indicated genotypes were analysed (bottom panels); error bars indicate standard deviation. **d**, Immunohistochemical staining for γ-H2AX and P-ATM$^{1981}$ in control and *Ah-cre* intestinal crypts and villi. The orange dotted line outlines the luminal side of the villus; the white dotted line outlines the outer side of the villus. DAPI (blue) was used for staining nuclei. **e**, Examination of γ-H2AX and P-ATM$^{1981}$ in cell extracts from control and *Ah-cre* intestinal crypts and villi by western blot assays. **f**, Sections of small intestines from β-NF-injected control and *Ah-cre* mice were processed for TUNEL (brown) and cleaved caspase 3 (red) assays. DAPI (blue) or haematoxylin was used for staining nuclei. Arrows point to TUNEL-positive or cleaved caspase-3-positive cells. Quantification of TUNEL and cleaved caspase-3-positive cells in crypts and villi is shown in the bottom panels. $n = 3$, 3 different animals with the indicated genotypes were analysed; error bars indicate standard deviation.

Close examination of haematoxylin-and-eosin-stained slides revealed increased numbers of pyknotic nuclei in *E2f1–3*-deficient crypts (data not shown). TdT-mediated dUTP nick end labelling (TUNEL) and cleaved caspase-3 assays confirmed the presence of apoptotic cells in crypts of *E2f1–3*-deficient intestines (Fig. 2f). We also observed increased p53 immunoreactivity in *E2f1–3*-deficient crypts (Supplementary Fig. 11a), which was reminiscent of previous work showing exquisite sensitivity of this cellular compartment to oncogene- and radiation-induced p53 responses[16]. Although p53 was

elevated in *E2f1–3*-deficient crypts, we failed to detect any significant increase in the expression of p53-responsive genes, and moreover, the conditional ablation of *p53* (β-NF treated *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*p53*$^{loxP/loxP}$) did not suppress the apoptosis caused by *E2f1–3* deficiency (Supplementary Fig. 11b, c). Together, these observations suggest that E2f1–3 are dispensable for cell division in the adult and that at least in the small intestine they function in a p53-independent manner to maintain DNA integrity and cell survival.

To understand the underlying mechanism for these unexpected results, we isolated crypt and villus cell populations from control and *E2f1–3*-deficient small intestines and analysed global gene expression profiles. Sample preparation and processing of the Affymetrix oligoarrays are described in the Methods section. We used an unbiased method similar to Gene Set Enrichment Analysis to identify genes that were differentially expressed[17]. Two variables contributed to the observed gene expression changes: cell compartment (crypt versus villus) and genotype (*E2f1–3*-deficient versus control). The cell compartment analysis compared gene expression in crypts and villi of the same genotype (Fig. 3a). For control small intestines, this revealed that among the ~45,000 genes queried, 1,207 genes were upregulated and 2,363 genes were downregulated as progenitor cells in the crypt migrated up into the villus and exited the cell cycle (>1.5-fold, $P < 0.0001$; Supplementary Fig. 12a and Supplementary Table 1). As expected, the expression of most known E2f targets, as defined by previous gene expression[18], reporter and chromatin immunoprecipitation assays[19] (Supplementary Fig. 12b), was markedly higher in control crypts than in associated villi (Fig. 3a, left panel), consistent with the proliferative status of crypts. For *E2f1–3*-deficient small intestines, the expression of E2f targets in crypts was only marginally higher than in their associated villi (Fig. 3a, right panel), suggesting that expression of these genes were either reduced in crypts, elevated in villi, or both.

The genotype analysis compared gene expression in *E2f1–3*-deficient versus control samples of the same cell compartment. This comparison revealed a modest but significant downregulation of E2f targets in progenitor cells of *E2f1–3*-deficient crypts, which included many but not all known classic targets such as *Cdc6*, *Ccna2*, *Ccne2*, *Top2* and *Hmgb2* (Fig. 3b, c, left panels and Supplementary Fig. 13a). We suspect that continued proliferation of *E2f1–3*-deficient progenitors in the small intestine when E2f targets are limiting probably contributes to replicative stress, DNA damage

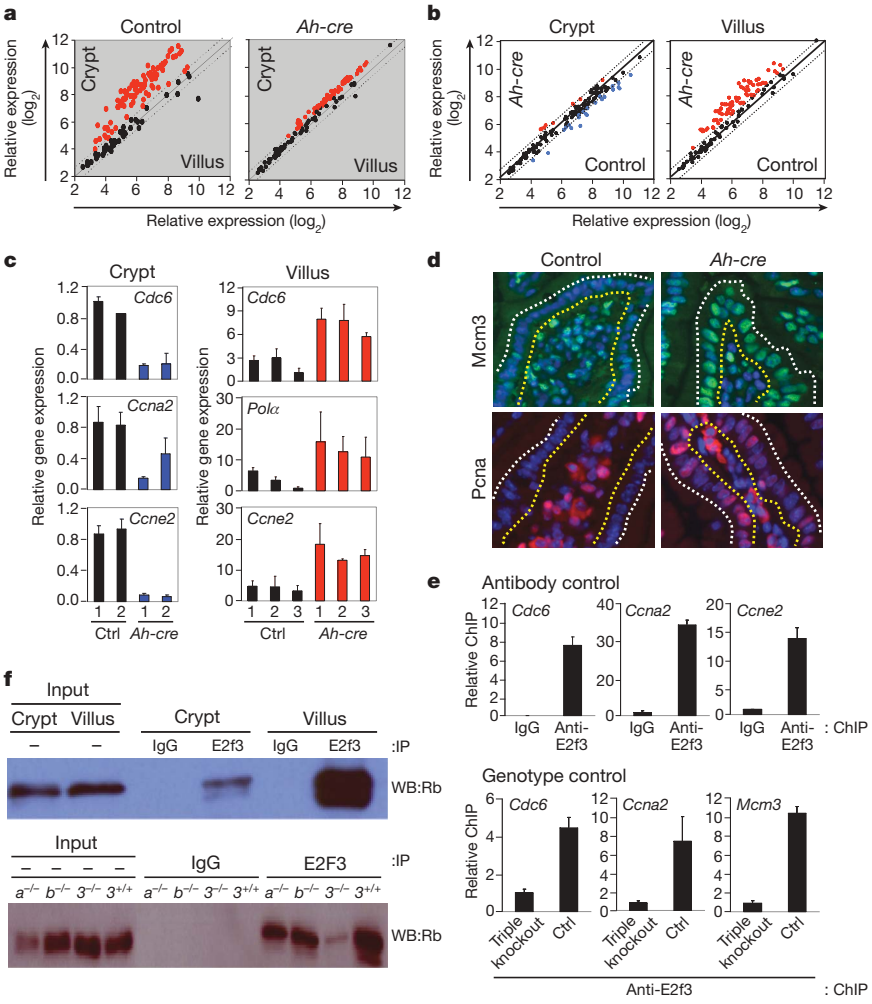



**Figure 3 | Repression of E2f target genes in *E2f1–3*-deficient villi.** **a**, Scatter plots comparing expression of known E2f target genes (see Supplementary Fig. 12b) between cell compartments (crypt and villus). Control, β-NF treated *E2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$; *Ahe-cre*, β-NF treated *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$. Genes with >1.5-fold increase in expression are depicted as red dots. **b**, Scatter plots comparing expression of known E2f target genes between genotypes (control and *Ah-cre* samples); $n = 3$ for each of the four samples. Red dots indicate genes whose expression increased >1.5-fold and blue dots indicate genes that decreased >1.5-fold. **c**, Quantitative real-time PCR was performed to compare the relative expression of selected E2f target genes in control and *Ah-cre* crypts (left panels) and villi (right panels) using specific primers (Supplementary Fig. 20). Error bars represent standard deviation from samples analysed in triplicate. **d**, Immunohistochemical staining of Mcm3 (green) and Pcna (red) in control and *Ah-cre* villi. DAPI (blue) was used for staining nuclei. Yellow dotted line outlines the luminal side of the villus; white dotted line outlines the outer side of the villus. Note that staining of blood cells in lumens of villi is nonspecific. **e**, The top panel shows chromatin immunoprecipitation (ChIP) assays using IgG or anti-E2f3 antibodies with lysates from wild-type villi (antibody control). The bottom panel shows ChIP assays using anti-E2f3 antibodies with lysates from wild type (ctrl) and *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ (triple knockout) villi (genotype control). Primers flanking known E2f-binding elements were used to detect the indicated gene promoters (Supplementary Fig. 19). Error bars represent standard deviation from samples analysed in triplicate. **f**, Co-immunoprecipitation assays of cell extracts prepared from control villi and crypts. Immunoprecipitations (IP) used anti-E2f3 antibody or IgG. Anti-Rb antibody was used to probe western blot (WB; top panel). The specificity of the anti-E2f3 antibody was evaluated in intestinal lysates derived from β-NF treated *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ ($3^{-/-}$), *E2f3a*$^{-/-}$ ($a^{-/-}$), *E2f3b*$^{-/-}$ ($b^{-/-}$) and *E2f3*$^{+/+}$ ($3^{+/+}$) mice. Anti-Rb antibody was used to probe western blot (WB; bottom panel).

and the observed increase in γ-H2AX labelling in these cells. Whether these aberrant processes are linked to the death of *E2f1–3*-deficient progenitor cells remains to be evaluated rigorously. The genotype comparison also revealed a remarkable upregulation of a large number of E2f targets in differentiated cells of the *E2f1–3*-deficient villus (Fig. 3b, c, right panels, and Supplementary Fig. 13a). Western blot assays and immunofluorescence staining showed that the accumulation of two of these E2f target gene products, Mcm3 and Pcna, was widespread throughout the *E2f1–3*-deficient villus (Fig. 3d and Supplementary Fig. 13b). Similarly, there was increased expression of E2f targets in differentiated *E2f1–3*-deficient cells of the retina and lens (Supplementary Fig. 13c; P.L.W., unpublished observations), suggesting a general role for E2f1–3 in transcriptional repression in post-mitotic cells *in vivo*. Chromatin immunoprecipitation (ChIP) assays using villus-enriched lysates derived from control and *E2f1–3*-deficient small intestines showed that E2f3 occupies E2f binding sites on classic E2f target promoters (Fig. 3e). Importantly, co-immunoprecipitation assays using intestinal epithelial cells derived from $E2f3^{-/-}$, $E2f3a^{-/-}$ and $E2f3b^{-/-}$ villi showed that both E2f3a/b isoforms[20–22] participate in a complex with the Rb protein (Fig. 3f). Consistent with this, Rb was found to be hypophosphorylated in the villus (Supplementary Fig. 14a). Together, these data indicate that E2f1–3 act as transcription activators in dividing progenitors, and as repressors (in complex with Rb) in differentiating cells of the small intestine.

The observation that E2f1–3 repress E2f targets and are dispensable for cell proliferation seems to contradict previous findings from the analysis of *Rb/E2f* double knockout animals[1,2,8]. Therefore, to explore thoroughly the mechanistic relationship between Rb and E2f1–3, we used the small intestine as an *in vivo* system where results could be uniformly compared across different genetic configurations. The *Ah-cre*-mediated inactivation of *Rb in utero* or in adult mice resulted in increased proliferation of cells in the villus compartment but not in the crypt (Fig. 4a, b and Supplementary Fig. 14b–e), indicating that *Rb*-deleted transit-amplifying cells failed to exit the cell cycle appropriately. There was, however, no concomitant increase in apoptosis or defect in cell differentiation (Supplementary Fig. 15a, b), and as a result, *Rb*-deficient villi appeared uniformly hyperplastic. The combined ablation of *E2f1*, *E2f2* and *E2f3* completely suppressed the unscheduled proliferation and hyperplasia caused by *Rb* deficiency (β-NF treated *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*Rb*$^{loxP/loxP}$ quadruple knockout; Fig. 4a, b and Supplementary Fig. 16a). Notably, the basal levels of proliferation in quadruple knockout crypts were indistinguishable from control or *E2f1–3*-deficient samples (Supplementary Fig. 16b), consistent with the rather normal development of *E2f1–3*-deficient small intestines containing an intact *Rb* gene.

The selective requirement for *E2f1–3* in the proliferation of *Rb*-deficient cells provided an opportunity to dissect possible cancer-specific mechanisms of E2f in cell cycle control. We therefore compared global gene expression programs in β-NF treated control, *Ah-creRb*$^{loxP/loxP}$ (*Rb*-deficient), *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ (*E2f1–3*-deficient) and *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*Rb*$^{loxP/loxP}$ (*Rb/E2f1–3*-deficient) intestinal epithelia. Several important insights came from this analysis. First, there were expansive gene expression differences between control and *Ah-creRb*$^{loxP/loxP}$ villi (1,290 upregulated and 487 downregulated genes; Fig. 4c and Supplementary Table 2), but relatively minor differences in their associated crypts (Supplementary Fig. 17a, b). Gene Ontology algorithms[23] identified a bias for differentially expressed genes involved in the regulation of transcription, DNA metabolic processes and cell cycle (Supplementary Table 3). Immunofluorescence and quantitative polymerase chain reaction with reverse transcription (RT–PCR) assays confirmed the marked accumulation of most E2f target genes in *Rb*-deficient villi (Supplementary Fig. 17c, d). From these data we conclude that *Rb* is critical for the repression of E2f targets at a time when progenitor cells commit to exit the cell cycle and terminally differentiate. Second, hierarchical clustering of all data sets showed that *E2f1–3*-deficient and *Rb/E2f1–3*-deficient tissues clustered together in a separate group from control and *Rb*-deficient tissues (Supplementary Fig. 18), indicating that some functions coordinated by E2f1–3 may be Rb-independent. Finally and most importantly, the expression levels of E2f targets in *E2f1–3*-deficient and *Rb/E2f1–3*-deficient villi were equivalent, and although higher than in control villi, they were substantially lower than in *Rb*-deficient villi (Fig. 4c). Quantitative RT–PCR assays confirmed the relative expression of E2f targets to be: control < *E2f1–3* deficient = *Rb/E2f1–3* deficient ≪ *Rb* deficient villi (Fig. 4d). From these data, we conclude that the supra-elevated expression of E2f targets observed in *Rb*-deficient villi is due to both 'derepression' (lacking intact Rb–E2f1–3 repressor complexes) and E2f1–3-mediated 'hyper-activation'. In the absence of E2f1–3-mediated hyper-activation, cells in *Rb/E2f1–3*-deficient villi fail to hyper-activate and thus do not accumulate sufficient levels of E2f targets to undergo 'ectopic' cell proliferation (this threshold level of expression is illustrated as a red dotted line in Fig. 4d).

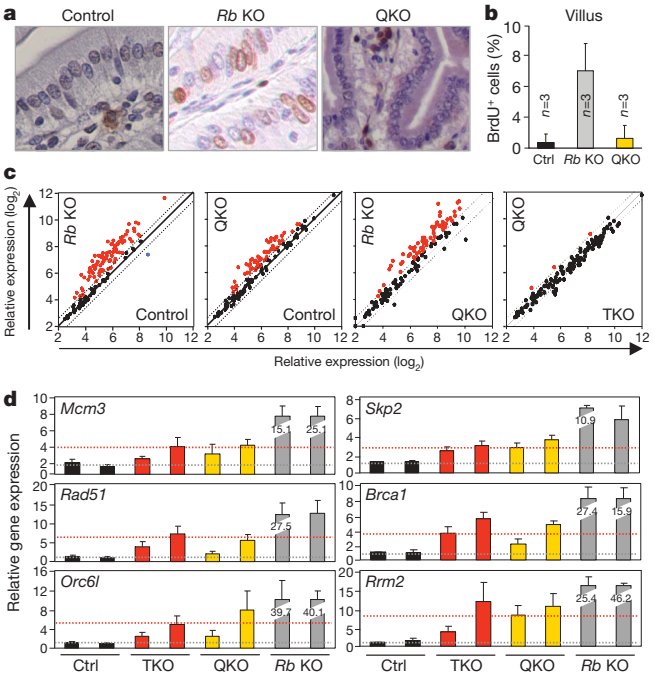



**Figure 4 | E2f1–3 contribute to the ectopic cell proliferation caused by *Rb*-deficiency.** **a**, BrdU analysis was performed in β-NF treated wild type (control), *Ah-creRb*$^{loxP/loxP}$ (*Rb* knockout (KO)) and *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*Rb*$^{loxP/loxP}$ (quadruple knockout (QKO)) small intestines. **b**, Quantification of BrdU incorporation. $n = 3$, 3 different animals with the indicated genotypes were analysed; error bars indicate standard deviation. **c**, Scatter plot analysis comparing differentially expressed E2f target genes in β-NF treated control, *Ah-creRb*$^{loxP/loxP}$ (*Rb* KO), *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ (TKO) and *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*Rb*$^{loxP/loxP}$ (QKO) villi; $n = 3$ for each of the eight samples. Red dots indicate genes whose expression increased >1.5-fold and blue dots indicate genes that decreased >1.5-fold. **d**, Quantitative RT–PCR analysis of selected E2f target genes in β-NF treated control (ctrl), *Ah-creRb*$^{loxP/loxP}$ (*Rb* KO), *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ (TKO) and *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$*Rb*$^{loxP/loxP}$ (QKO) villi. The normal basal level of E2f target expression is illustrated as a grey dotted line and the threshold level of E2f target expression required for ectopic proliferation is illustrated as a red dotted line. Error bars represent standard deviation from samples analysed in triplicate.

We provide overwhelming evidence showing that normal cell proliferation in mice can be maintained in the absence of activator E2fs. We conclude that E2f1–3, like G1 Cdks[24–26], are not as critical for normal cell proliferation in mammals as original studies implied[3,4,27–30]. However, not all is well in the absence of E2f1–3, as β-NF treated *Ah-creE2f1*$^{-/-}$*E2f2*$^{-/-}$*E2f3*$^{loxP/loxP}$ dividing progenitors in the small intestine undergo apoptosis. A pro-survival role for E2f1–3 was also evident in retinal progenitor cells of the mouse[15]; however, in the retina cell death was p53 dependent whereas in the small intestine it was p53

independent. Thus, the sensitivity of *E2f1-3*-deficient cells to p53 activation varies considerably across tissue types.

The findings presented here also expose dual functions for E2f1–3 in transcription activation and repression *in vivo*. In dividing progenitor cells, when Rb is inactive (hyperphosphorylated), free E2f1–3 are used to optimally activate the expression of target genes. The inability to do so in *E2f1–3*-deficient tissues still permits cells to replicate their DNA and divide, but at the cost of increased DNA damage and cell death. As cells commit to a terminally differentiated fate, phosphorylated Rb is dephosphorylated and forms a physical complex with E2f1–3 proteins. We propose that this is not just to sequester E2f activators but rather, to form the first repressive complex that is necessary to downregulate E2f targets and usher transit-amplifying cells out of the cell cycle. Once cells exit the cell cycle, other E2F repressor complexes accumulate, including p130–E2f4 and p107–E2f4, to more permanently enforce the repression of E2f targets. Given that inactivation of *Rb*, but not *p107* or *p130* (ref. 31), induces ectopic cell divisions in the small intestine, we suggest that Rb has a unique role in transit-amplifying cells that is dependent on its ability to associate with E2f1–3. Maintenance of quiescence in terminally differentiated cells of the villus, however, is a function that is shared among all members of the Rb family[31]. This work challenges the current paradigm of cell cycle control and provides a unified molecular view of how the dual functions of E2f1–3 in transcriptional activation and repression are used *in vivo* to control normal versus *Rb*-mutant or cancer cell cycles.

## METHODS SUMMARY

Mice (*E2f1*$^{-/-}$, *E2f2*$^{-/-}$, *E2f3*$^{f/f}$, *Ah-cre* and *Rb*$^{f/f}$) used for the studies were in a mixed background (129SvEv, C57BL/6NTac and FVB/NTac). β-Napthoflavone (Sigma; N3633-5G) was administered into 2-month-old *Ah-cre* mice three times within 24 h as described previously[14], and mice were harvested 7 or 90 days later. β-Napthoflavone was also injected into pregnant female mice at 15.5 days post coitum for analysis of embryos at E18.5. Villus and crypt fractions were isolated as previously described[8]. Three independent samples from each genetic group were used for gene expression analysis by Affymetrix microarray. Analysis of gene expression data was performed using BRB-array tools developed by R. Simon and A. Peng Lam of the National Cancer Institute. Gene Ontologies were predicted by DAVID (Database for Annotation, Visualization and Integrated Discovery) Bioinformatics Resources at the National Institute of Allergy and Infectious diseases, NIH. X-gal staining, real-time RT–PCR, 5-bromodeoxyuridine (BrdU), ChIP and TUNEL assays were performed as previously described[8,20]. Primers for ChIP, real-time RT–PCR and genotyping are listed in Supplementary Fig. 19a, b. Antibodies used for western blot or immunohistochemical staining are listed in Supplementary Fig. 19c.

**Received 2 September; accepted 17 November 2009.**


1. Iaquinta, P. J. & Lees, J. A. Life and death decisions by the E2F transcription factors. *Curr. Opin. Cell Biol.* **19,** 649–657 (2007).
2. Dimova, D. K. & Dyson, N. J. The E2F transcriptional network: old acquaintances with new faces. *Oncogene* **24,** 2810–2826 (2005).
3. DeGregori, J., Leone, G., Miron, A., Jakoi, L. & Nevins, J. R. Distinct roles for E2F proteins in cell growth control and apoptosis. *Proc. Natl Acad. Sci. USA* **94,** 7245–7250 (1997).
4. Johnson, D. G., Schwarz, J. K., Cress, W. D. & Nevins, J. R. Expression of transcription factor E2F1 induces quiescent cells to enter S phase. *Nature* **365,** 349–352 (1993).
5. Wu, L. *et al.* The E2F1–3 transcription factors are essential for cellular proliferation. *Nature* **414,** 457–462 (2001).
6. Timmers, C. *et al.* E2f1, E2f2, and E2f3 control E2F target expression and cellular proliferation via a p53-dependent negative feedback loop. *Mol. Cell. Biol.* **27,** 65–78 (2007).
7. Sharma, N. *et al.* Control of the p53-p21$^{CIP1}$ axis by *E2f1*, *E2f2*, and *E2f3* is essential for $G_1$/S progression and cellular transformation. *J. Biol. Chem.* **281,** 36124–36131 (2006).
8. Saenz-Robles, M. T. *et al.* Intestinal hyperplasia induced by simian virus 40 large tumor antigen requires E2F2. *J. Virol.* **81,** 13191–13199 (2007).
9. Rowland, B. D. & Bernards, R. Re-evaluating cell-cycle regulation by E2Fs. *Cell* **127,** 871–874 (2006).
10. Murga, M. *et al.* Mutation of E2F2 in mice causes enhanced T lymphocyte proliferation, leading to the development of autoimmunity. *Immunity* **15,** 959–970 (2001).
11. Iglesias, A. *et al.* Diabetes and exocrine pancreatic insufficiency in E2F1/E2F2 double-mutant mice. *J. Clin. Invest.* **113,** 1398–1407 (2004).
12. Cloud, J. E. *et al.* Mutant mouse models reveal the relative roles of E2F1 and E2F3 *in vivo*. *Mol. Cell. Biol.* **22,** 2663–2672 (2002).
13. van der Flier, L. G. & Clevers, H. Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu. Rev. Physiol.* **71,** 241–260 (2009).
14. Ireland, H. *et al.* Inducible Cre-mediated control of gene expression in the murine gastrointestinal tract: effect of loss of β-catenin. *Gastroenterology* **126,** 1236–1246 (2004).
15. Chen, D. *et al.* Division and apoptosis of E2f-deficient retinal progenitors. *Nature* doi:10.1038/nature08544 (this issue).
16. Coopersmith, C. M. & Gordon, J. I. γ-Ray-induced apoptosis in transgenic mice with proliferative abnormalities in their intestinal epithelium: re-entry of villus enterocytes into the cell cycle does not affect their radioresistance but enhances the radiosensitivity of the crypt by inducing p53. *Oncogene* **15,** 131–141 (1997).
17. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34,** 267–273 (2003).
18. Kong, L. J., Chang, J. T., Bild, A. H. & Nevins, J. R. Compensation and specificity of function within the E2F family. *Oncogene* **26,** 321–327 (2007).
19. Xu, X. *et al.* A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res.* **17,** 1550–1561 (2007).
20. Chong, J. L. *et al.* E2f3a and E2f3b contribute to the control of cell proliferation and mouse development. *Mol. Cell. Biol.* **29,** 414–424 (2009).
21. Leone, G. *et al.* E2F3 activity is regulated during the cell cycle and is required for the induction of S phase. *Genes Dev.* **12,** 2120–2130 (1998).
22. Leone, G. *et al.* Identification of a novel E2F3 product suggests a mechanism for determining specificity of repression by Rb proteins. *Mol. Cell. Biol.* **20,** 3626–3632 (2000).
23. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
24. Malumbres, M. & Barbacid, M. Mammalian cyclin-dependent kinases. *Trends Biochem. Sci.* **30,** 630–641 (2005).
25. Martín, A. *et al.* Cdk2 is dispensable for cell cycle inhibition and tumor suppression mediated by p27$^{Kip1}$ and p21$^{Cip1}$. *Cancer Cell* **7,** 591–598 (2005).
26. Malumbres, M. *et al.* Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell* **118,** 493–504 (2004).
27. Russell, P. & Nurse, P. *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*: a look at yeasts divided. *Cell* **45,** 781–782 (1986).
28. Helin, K. *et al.* A cDNA encoding a pRB-binding protein with properties of the transcription factor E2F. *Cell* **70,** 337–350 (1992).
29. Kaelin, W. G. Jr *et al.* Expression cloning of a cDNA encoding a retinoblastoma-binding protein with E2F-like properties. *Cell* **70,** 351–364 (1992).
30. Nevins, J. R. Transcriptional regulation. A closer look at E2F. *Nature* **358,** 375–376 (1992).
31. Haigis, K., Sage, J., Glickman, J., Shafer, S. & Jacks, T. The related retinoblastoma (pRb) and p130 proteins cooperate to regulate homeostasis in the intestinal epithelium. *J. Biol. Chem.* **281,** 638–647 (2006).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank L. Rawahneh, J. Moffitt and R. Rajmohan for technical assistance with histology. We also thank A. de Bruin and S. Naidu for assistance in analysing histological slides. We are thankful to J. Groden, A. Simcox and D. Guttridge for their critical comments. This work was funded by NIH grants to G.L. (R01CA85619, R01CA82259, R01HD04470, P01CA097189) and NIH grant to J.M.P. (CA098956); J.-L.C. is the recipient of a DoD award (BC061730). P.L.W. was supported by NIH training grant 5 T32 CA106196-04.

**Author Contributions** M.L.R., J.M.P. and G.L. designed and supervised this study, analysed data, and helped write and edit the manuscript. J.-L.C., P.L.W. and M.T.S.-R. designed and performed experiments, collected and analysed data, and co-wrote the paper. V.N., A.F., Y.M.G., N.S., H.-Z.C., M.O., S.-H.W., P.T., B.C. and L.M. technically assisted with experiments and collected and analysed data. D.C. and R.B. performed and analysed gene expression of retina. J.P.H. and P.G.C. contributed to the analysis and comparison of gene microarray data. D.J.W. and O.J.S. contributed to the generation of key reagents.

**Author Information** All microarray data have been deposited at the Gene Expression Omnibus at the National Center for Biotechnology Information under accession number GSE16454. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.L. (Gustavo.Leone@osumc.edu).

LETTERS

# Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks

Yaron Galanty[1], Rimma Belotserkovskaya[1], Julia Coates[1], Sophie Polo[1], Kyle M. Miller[1] & Stephen P. Jackson[1]

**DNA double-strand breaks (DSBs) are highly cytotoxic lesions that are generated by ionizing radiation and various DNA-damaging chemicals. Following DSB formation, cells activate the DNA-damage response (DDR) protein kinases ATM, ATR and DNA-PK (also known as PRKDC). These then trigger histone H2AX (also known as H2AFX) phosphorylation and the accumulation of proteins such as MDC1, 53BP1 (also known as TP53BP1), BRCA1, CtIP (also known as RBBP8), RNF8 and RNF168/RIDDLIN into ionizing radiation-induced foci (IRIF) that amplify DSB signalling and promote DSB repair[1,2]. Attachment of small ubiquitin-related modifier (SUMO) to target proteins controls diverse cellular functions[3–6]. Here, we show that SUMO1, SUMO2 and SUMO3 accumulate at DSB sites in mammalian cells, with SUMO1 and SUMO2/3 accrual requiring the E3 ligase enzymes PIAS4 and PIAS1. We also establish that PIAS1 and PIAS4 are recruited to damage sites via mechanisms requiring their SAP domains, and are needed for the productive association of 53BP1, BRCA1 and RNF168 with such regions. Furthermore, we show that PIAS1 and PIAS4 promote DSB repair and confer ionizing radiation resistance. Finally, we establish that PIAS1 and PIAS4 are required for effective ubiquitin-adduct formation mediated by RNF8, RNF168 and BRCA1 at sites of DNA damage[7–11]. These findings thus identify PIAS1 and PIAS4 as components of the DDR and reveal how protein recruitment to DSB sites is controlled by coordinated SUMOylation and ubiquitylation.**

Mammalian cells express SUMO1 and the highly-related proteins SUMO2 and SUMO3 (SUMO2/3). These somewhat functionally-redundant proteins[12] are structurally related to ubiquitin and are covalently attached to target proteins by a SUMO-conjugation system consisting of an E1 activating enzyme (SAE1/SAE2), an E2 ligase (UBC9, also known as UBE2I) and various E3 ligases with differing target-protein specificities[3,4]. Involvement of the SUMO pathway in aspects of the DDR was previously reported (for review, see ref. 5). Notably, we found that, whereas SUMO1 exhibited pan-nuclear staining in untreated human cells, four hours after ionizing radiation treatment, it formed nuclear foci that largely co-localized with 53BP1, suggesting them to be IRIF (Fig. 1a). Similarly, transfected haemagglutinin (HA)-epitope-tagged SUMO1 and SUMO3 formed IRIF (Fig. 1a; SUMO2/3 foci that do not co-localize with 53BP1 presumably reflect SUMO conjugates in other structures, including PML bodies). Next, we employed laser micro-irradiation to induce DNA-damage tracts (laser-lines) in living cells[13,14]. This showed that endogenous SUMO1 and SUMO2/3 (the antibody does not discriminate between these), together with HA–SUMO1 and HA–SUMO3, accumulated in laser-lines (Fig. 1b). Moreover, live imaging of cells containing green-fluorescent-protein (GFP)-tagged 53BP1 or red-fluorescent-protein (RFP)-tagged SUMO1, SUMO2 or SUMO3 showed that all exhibited similar recruitment kinetics: accrual being detectable five minutes after micro-irradiation, peaking in intensity at two to four hours

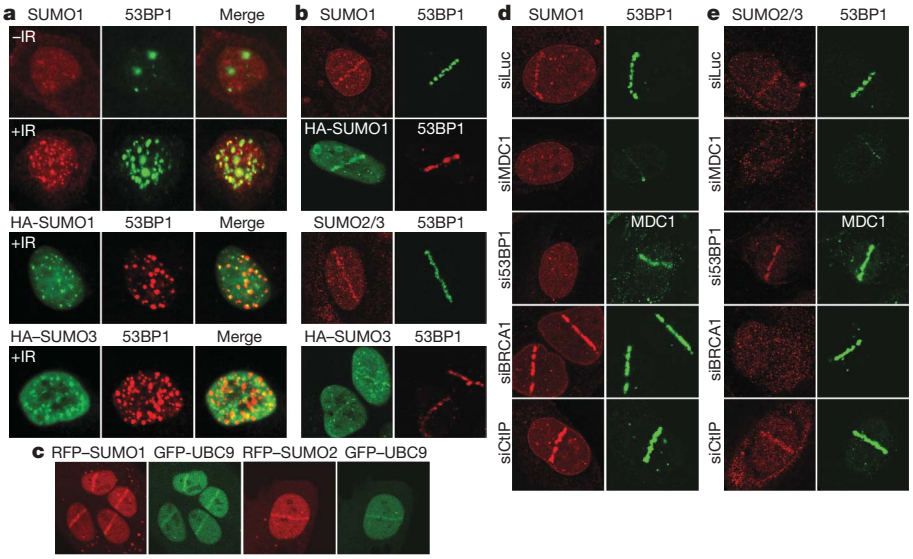


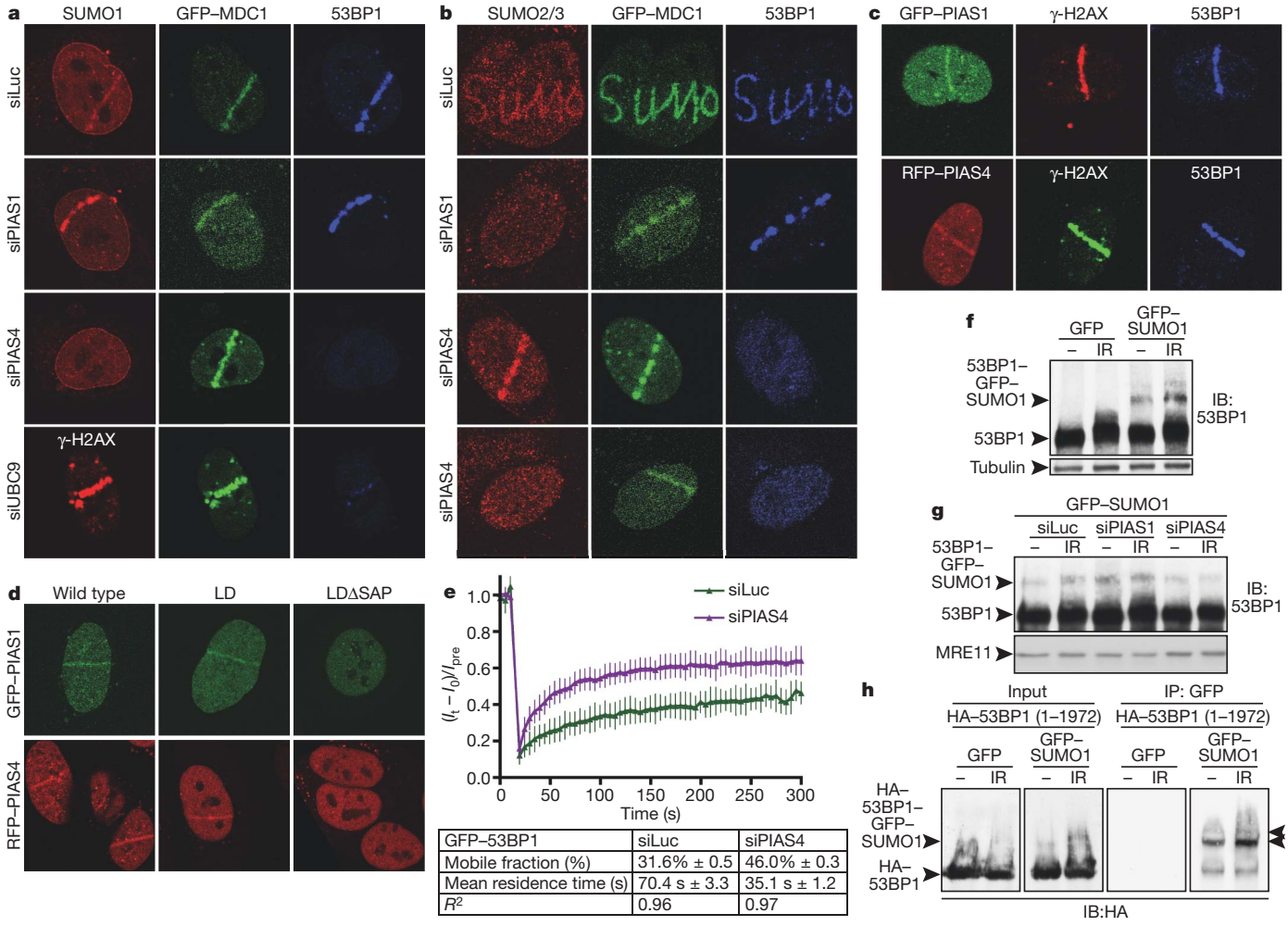

**Figure 1 | SUMOs and UBC9 accumulate at DNA-damage sites by mechanisms requiring MDC1, 53BP1 and BRCA1. a**, U2OS cells or U2OS cells transfected with HA–SUMO1 or HA–SUMO3 were irradiated (5 Gy; +IR) or mock-irradiated (−IR) and probed. **b**, As in **a** but with laser micro-irradiation. **c**, U2OS cells co-transfected with GFP–UBC9 and RFP–SUMO1 or RFP–SUMO2 were micro-irradiated and live cells imaged after 20 min. **d**, **e**, U2OS cells transfected with siRNAs were laser micro-irradiated and probed. For siRNA depletions, see Fig. 3e and Supplementary Fig. 10a.

[1]The Wellcome Trust and Cancer Research UK Gurdon Institute, and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK.

and then gradually diminishing (Supplementary Figs 1a–c, 2a and b). Furthermore, we observed SUMO1 and SUMO2/3 accumulation with varying intensities in both $G_1$ and $S/G_2$ cells (Supplementary Fig. 2c). Consistent with SUMOylation actively occurring at damage sites, UBC9 (the only known SUMO E2) accumulated at damaged regions with similar kinetics to SUMO (Fig. 1c and Supplementary Figs 1b, 1d and 2d). Furthermore, we observed faint recruitment of the SUMO E1 component, SAE1, to laser-lines (data not shown), in accord with SAE1 recently being identified as a potential ATM/ATR target[15].

In line with SUMO accumulation in IRIF and laser-lines representing responses to DSBs, such accumulation was reduced when cells were pre-incubated with KU-55933, a specific ATM inhibitor[16] (Supplementary Fig. 3a), whereas accumulation of SUMO1, and to a lesser extent SUMO2/3, was enhanced by depletion of CtIP or MMS21 (also known as NSMCE2), which promote DNA repair[17,18] (Figs 1d, e and Supplementary Figs 4a, b; see Fig. 3e for CtIP depletion and Supplementary Fig. 10 for other depletions). Furthermore, we observed markedly reduced SUMO1 and SUMO2/3 accumulation at damaged sites in cells that were defective in RNF168 or had been treated with short-interfering RNAs (siRNAs) to deplete MDC1 or RNF8 (Figs 1d and 1e, and Supplementary Figs 3b and 3c). Because MDC1, RNF8 and RNF168 control the retention of 53BP1 and BRCA1 at DNA-damage sites[7–9,11,19–23], we tested whether depleting these factors affected SUMO accrual. Indeed, 53BP1 depletion impaired SUMO1 but not SUMO2/3 accumulation in laser-lines (Figs 1d and 1e). Conversely, BRCA1 depletion abolished SUMO2/3 but not SUMO1 accrual (Figs 1d and 1e). Collectively, these data suggested that DNA damage is channelled into 53BP1-SUMO1 or BRCA1-SUMO2/3 pathways.

The different accumulation requirements for SUMO1 and SUMO2/3 suggested that their conjugation might require different E3 ligases. By siRNA depletion of various SUMO E3 ligases (Supplementary Fig. 10), we found that most were not required for SUMO1 or SUMO2/3 accrual at DNA damage sites (Supplementary Figs 4a and 4b). Strikingly, however, depletion of the PIAS4 E3 ligase markedly reduced SUMO1 accrual on laser-lines (Figs 2a and 2b; note that MDC1 recruitment still occurred). Nevertheless, in certain cells, PIAS4 depletion also impaired SUMO2/3 (and 53BP1) accumulation (Fig. 2b, bottom panels), indicating that PIAS4 controls both SUMO1 and SUMO2/3 accrual. Accordingly, PIAS4 depletion impaired the accumulation of GFP–SUMO3 at laser-lines (data not shown). In parallel, we found that PIAS1 depletion markedly reduced SUMO2/3 accumulation at sites of DNA damage in all cells, but did not affect SUMO1 accrual (Figs 2a, 2b and Supplementary Figs 4a and 4b). Supporting a model in which PIAS4 and PIAS1 mediate SUMO conjugation at DSB sites, RFP-tagged PIAS4 and GFP-tagged PIAS1 were



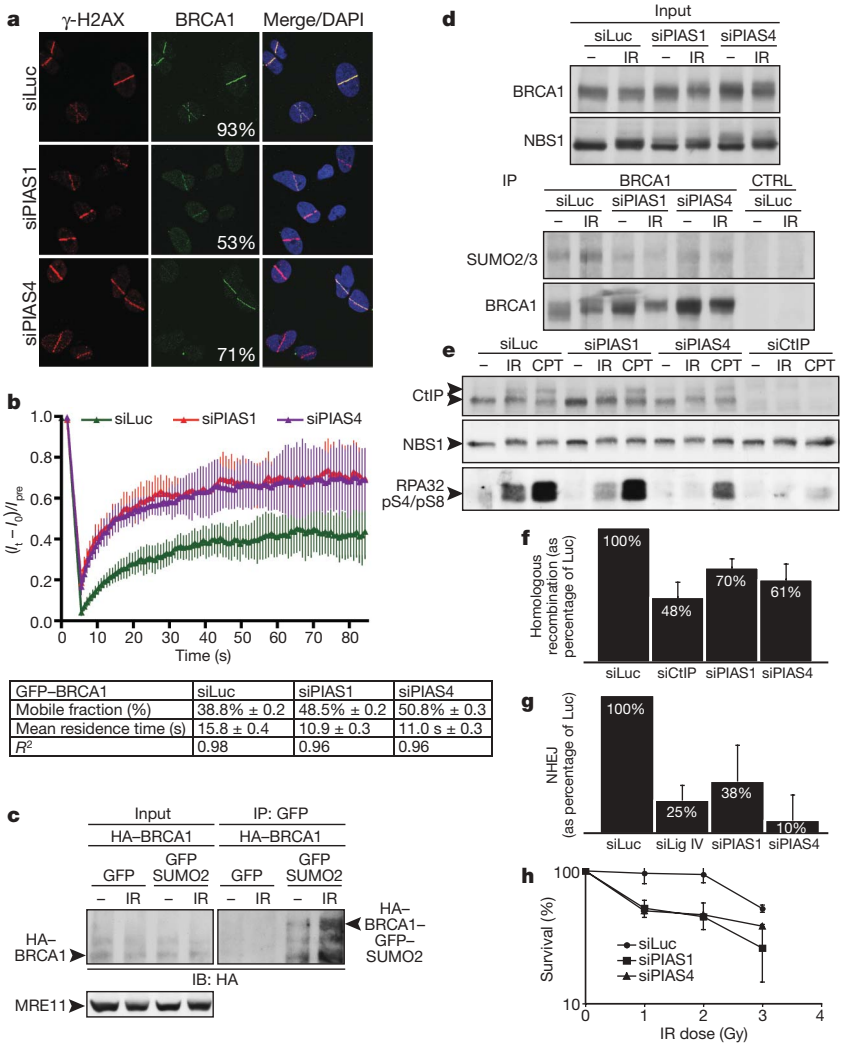

| GFP–53BP1 | siLuc | siPIAS4 |
|---|---|---|
| Mobile fraction (%) | 31.6% ± 0.5 | 46.0% ± 0.3 |
| Mean residence time (s) | 70.4 s ± 3.3 | 35.1 s ± 1.2 |
| $R^2$ | 0.96 | 0.97 |

**Figure 2 | PIAS1 and PIAS4 are recruited to DNA-damage sites and mediate 53BP1 recruitment and SUMOylation.** **a**, **b**, U2OS cells stably expressing GFP–MDC1 were treated, micro-irradiated and probed (Supplementary Figs 4c and 10a–c for quantifications and depletions, respectively). **c**, Cells expressing GFP–PIAS1 or RFP–PIAS4 were micro-irradiated and probed. **d**, Cells expressing GFP–PIAS1 or RFP–PIAS4 wild-type, ligase dead (LD), delta SAP (data not shown) or ligase-dead delta SAP (LDΔSAP) were micro-irradiated and imaged. **e**, Cells stably expressing GFP–53BP1 subjected to FRAP ($n = 11$ independent measurements; error bars = s.d.). **f**–**h**, U2OS cells stably expressing GFP–SUMO1 or GFP, MRE11 and Tubulin used as loading control (**f**, **g**), or HEK293 cells co-transfected with full-length (1–1972) HA–53BP1 and GFP–SUMO1 or GFP (**h**) were treated with or without IR (10 Gy). IB, Immunoblotting.

recruited to laser-lines with similar kinetics to SUMO and UBC9 (Fig. 2c and 2d). Furthermore, for both PIAS4 and PIAS1, recruitment required their N-terminal SAP domain—originally defined as a DNA/RNA binding motif[24]—but was not impaired by mutations predicted to abolish their SUMO E3-ligase functions (Fig. 2d). When expressed alone, however, the SAP domains of PIAS1 and PIAS4 were not detectably recruited to laser-lines, revealing that additional parts of these proteins are required for effective recruitment (data not shown).

Strikingly, PIAS4 depletion by either of two independent siRNA oligonucleotides, but not depletion of any other E3 enzyme tested, severely impaired 53BP1 accumulation in laser-lines and in IRIF (Figs 2a, b and Supplementary Figs 4a–d; demonstration of siRNA specificity is provided by use of a point-mutated siRNA-resistant PIAS4 construct in Supplementary Figs 5a and b). In accord with this, UBC9 depletion impaired 53BP1 accumulation, while histone H2AX phosphorylation (γ-H2AX) and MDC1 recruitment still ensued (Fig. 2a). Furthermore, fluorescence-recovery after photo-bleaching (FRAP) assays established that PIAS4 depletion significantly reduced the residence time of 53BP1 in laser-lines and increased the mobile fraction of 53BP1 molecules in these locations (Fig. 2e; representative images are shown in Supplementary Fig. 5d). By contrast, RFP–PIAS4 accrual in laser lines was not impaired by 53BP1 depletion (Supplementary Fig. 5c), implying that PIAS4 acts upstream of 53BP1.

During the above studies, we noted that the ionizing radiation-induced shift in 53BP1 electrophoretic mobility on SDS-polyacrylamide gels was reduced by PIAS4 depletion (data not shown). Consistent with this mobility shift at least in part reflecting 53BP1 SUMOylation, the migration of endogenously expressed 53BP1 on SDS-PAGE was shifted yet further in cells expressing GFP-tagged SUMO1 (Fig. 2f); furthermore, this shift was diminished by PIAS4 depletion but not by PIAS1 depletion (Fig. 2g). To test directly for 53BP1 SUMOylation, we transiently co-expressed HA-tagged 53BP1 with GFP–SUMO1 or GFP in cells, then performed GFP immunoprecipitations. Western immunoblotting of resulting samples with an anti-HA antibody established that 53BP1 was indeed SUMOylated in an ionizing radiation-inducible manner (Fig. 2h), a conclusion supported by reciprocal immunoprecipitation-western experiments (Supplementary Fig. 6a) and by experiments with endogenous 53BP1 and SUMO1 (Supplementary Fig. 6b; this also showed that 53BP1 SUMOylation was reduced by depleting PIAS4 but not PIAS1). Studies with cells expressing 53BP1 truncations revealed that both the amino terminal (residues 1–1052) and carboxy terminal (1052–1972) regions of 53BP1




| GFP–BRCA1 | siLuc | siPIAS1 | siPIAS4 |
|---|---|---|---|
| Mobile fraction (%) | 38.8% ± 0.2 | 48.5% ± 0.2 | 50.8% ± 0.3 |
| Mean residence time (s) | 15.8 ± 0.4 | 10.9 ± 0.3 | 11.0 s ± 0.3 |
| $R^2$ | 0.98 | 0.96 | 0.96 |

**Figure 3 | PIAS1 and PIAS4 promote BRCA1 accumulation and SUMOylation, RPA phosphorylation, and DSB repair. a**, U2OS cells treated, micro-irradiated and probed as indicated; representative images with % of γ-H2AX positive cells also positive for BRCA1, each image represents >200 γ-H2AX-positive cells in two independent experiments. **b**, U2OS cells stably expressing GFP–BRCA1 and Flag–BARD1 were subjected to FRAP; data from luciferase (Luc, $n = 7$ independent measurements), PIAS1 ($n = 8$) and PIAS4 ($n = 11$); error bars = s.d. **c**, Essentially as Fig. 2h, except cells were co-transfected with GFP–SUMO1, HA–BRCA1 and Flag–BARD1. IP, Immunoprecipitation. MRE11 (also known as MRE11A) was used as loading control. **d**, **e**, Extracts were prepared and analysed 2 h following mock (-) or 10 Gy ionizing radiation treatment. NBS1 (also known as NBN) was used as loading control. **f**–**h**, Effects of PIAS1/4 depletion on homologous recombination-mediated gene-conversion (**f**), NHEJ (**g**) and ionizing radiation sensitivity (**h**); error bars = s.d.; data accumulated over four independent experiments (in each of **f**–**h**).

can be SUMOylated and suggested that C-terminal SUMOylation occurs between residues 1052 and 1710 (Supplementary Figs 6c–e). While these data indicated that DNA-damage-induced 53BP1 SUMOylation occurs, we note that this cannot account for all the PIAS-dependent SUMOylation signals observed in IRIF or laser-lines. Consequently, there must be additional DDR factors (some of which might have been identified in previous studies[5]) that are targeted for DNA-damage induced, PIAS-mediated SUMOylation.

In parallel work, we found that both PIAS1 or PIAS4 depletion reduced the proportion of damaged (γ-H2AX-positive) cells displaying BRCA1 accumulation and decreased BRCA1 staining intensity in those cells still exhibiting BRCA1 accrual (Fig. 3a and Supplementary Fig. 7b; cells with weak BRCA1 staining in Fig. 3a were counted positive). By employing cDNA complementation studies, we established that BRCA1 accrual required the SAP domain and E3-ligase activity of PIAS1 (Supplementary Figs 7c–e). Furthermore, FRAP analyses revealed that PIAS1 or PIAS4 depletion reduced the residence time of GFP–BRCA1 at damaged sites and increased the mobile fraction of BRCA1 molecules (Fig. 3b; see representative images in Supplementary Fig. 7b). Through using epitope-tagged SUMO2 and BRCA1 in immunoprecipitation-western studies, we also established that BRCA1 is SUMOylated and that this is enhanced upon ionizing radiation treatment (Fig. 3c). Accordingly, probing western blots of BRCA1 immunoprecipitates for SUMO2/3 revealed that ionizing radiation enhanced BRCA1 SUMOylation in a manner promoted by both PIAS1 and PIAS4 (Fig. 3d).

DSBs can be processed into single-stranded DNA that is bound by replication protein A (RPA) to promote ATR signalling and DSB repair by homologous recombination[17]. Notably, RPA accumulation in laser-lines (whether normalized or not to cell cycle profiles in Supplementary Fig. 10e) was impaired by PIAS1 or PIAS4 depletion (Supplementary Figs 8a–c). Furthermore, phosphorylation of the 34 kDa subunit of RPA on Ser 4 and Ser 8 (pS4/pS8) in response to ionizing radiation or camptothecin treatment was diminished by PIAS4 depletion, whereas PIAS1 depletion impaired ionizing radiation-induced but not camptothecin-induced RPA phosphorylation (Fig. 3e; CtIP depletion also impaired RPA phosphorylation, as previously reported[17]). Consistent with these findings and the involvement of BRCA1 and RPA in DNA repair by homologous recombination[17,25,26], PIAS1 or PIAS4 depletion reduced homologous recombination in a cell-based gene conversion assay[27] (Fig. 3f). PIAS1 and PIAS4 depletion also impaired DSB repair by non-homologous end-joining (NHEJ) as assessed by a cell-based plasmid-integration assay[28] (Fig. 3g) and resulted in ionizing radiation hypersensitivity (Fig. 3h).

Accumulation of 53BP1, BRCA1 and ubiquitin conjugates at DSB sites requires the ubiquitin E3 ligases, RNF8 and RNF168, which ubiquitylate histones H2A and H2AX[7–9,11,22]. Furthermore, it has been reported that in both *Caenorhabditis elegans* and mammalian cells, ubiquitin-conjugate formation at DNA-damage sites requires BRCA1 E3-ubiquitin ligase activity[10,29], although other groups have reported the effect of BRCA1 depletion on ubiquitin accrual to be only minor[7,9,11]. In our assays, we found that, as for BRCA1 depletion, PIAS1 or PIAS4 depletion dramatically impaired ubiquitin-conjugate accumulation (as detected by the FK2 antibody) in laser-lines, while GFP–RNF8 accumulation appeared normal (Figs 4a, b and Supplementary Fig. 9a). Furthermore, PIAS4 depletion but not PIAS1 depletion markedly impaired histone H2A ubiquitylation at damaged sites (Fig. 4c and Supplementary Fig. 9b). Consistent with PIAS4 being required for DNA-damage-induced accrual of 53BP1, BRCA1, FK2–ubiquitin conjugates and ubiquitin–H2A, the recruitment of endogenous RNF168 to damaged regions was impaired in PIAS4 depleted cells (Fig. 4d and Supplementary Figs 9c and 9d). By contrast, RNF168 still assembled at damage sites in PIAS1-depleted cells (Fig. 4d; as shown previously[7,11], RNF168 accrual was RNF8 dependent). Because 53BP1 still accumulated under conditions where the FK2–ubiquitin signal was severely impaired (upon BRCA1 or PIAS1 depletion; Figs 1d, 1e, 2a, 2b and 4a), these data implied that 53BP1 recruitment does not require ubiquitin conjugates recognized by the FK2 antibody but, instead, relies on other ubiquitylated proteins (most likely H2A and H2AX). Significantly, depletion of RNF8 or RNF168, although abolishing 53BP1 accrual at sites of DNA damage, did not affect accumulation of GFP–PIAS1 or RFP–PIAS4 (Figs 4e and 4f). We therefore conclude that PIAS1 and PIAS4 function in parallel with RNF8 to orchestrate RNF8-, RNF168- and BRCA1-dependent accumulation of ubiquitin conjugates at DNA-damage sites. Only PIAS4, however, is needed for RNF8- and RNF168-mediated H2A and possibly H2AX ubiquitylation.

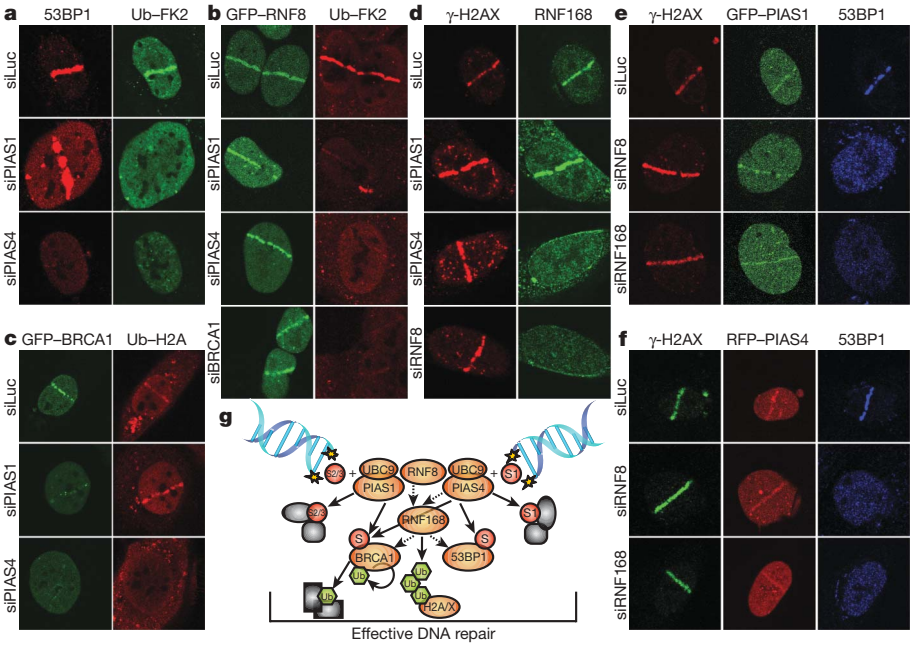



**Figure 4 | Linkage between PIAS1/4 and RNF8/168.** **a**, U2OS cells were treated and probed as indicated. **b**, **c**, As (**a**) except cells stably expressed GFP–RNF8 or GFP–BRCA1/Flag–BARD1; see Supplementary Figs 9a and b for quantifications. **d**, U2OS cells were treated and probed as indicated; see Supplementary Figs 9d and 10c for quantifications and siRNA efficiencies, respectively. **e**, **f**, U2OS cells stably expressing GFP–PIAS1 (**e**) or RFP–PIAS4 (**f**) were treated and probed as indicated. **g**, Model; dashed arrows indicate protein requirements for accumulation, solid arrows indicate target-protein modifications. Ub, ubiquitin. S, SUMO.

Our findings invoke a model in which PIAS1 and PIAS4 act in parallel but overlapping SUMO-conjugation pathways to control the DDR (Fig. 4g). In this regard, we note that mouse knockout studies have revealed that PIAS1 or PIAS4 loss is tolerated, whereas deletion of both leads to embryonic lethality and an inability to derive viable cells[30]. Significantly, whereas both PIAS1 and PIAS4 promote FK2–ubiquitin-adduct accumulation, only PIAS4 is needed for accrual of RNF168 and ubiquitylated H2A at DNA-damage sites. An attractive explanation for these and other data is that, after being recruited by RNF8-, PIAS1- and PIAS4-dependent mechanisms, BRCA1 (together with BARD1) is itself the major ubiquitin E3 ligase for generating FK2-reactive ubiquitin conjugates. Significantly, after PIAS1 or PIAS4 depletion, we still detect weak association of BRCA1 at damage sites but not the accumulation of BRCA1-dependent FK2–ubiquitin conjugates. Consequently, we speculate that PIAS1- and PIAS4-dependent SUMOylation of BRCA1—and in all likelihood various other DDR proteins—not only mediates the stable association of BRCA1 with DNA-damage sites but also promote BRCA1 ubiquitin-ligase activity. Furthermore, we found that GFP–RNF8 recruitment still occurred upon PIAS1 or PIAS4 depletion, showing that RNF8 recruitment is insufficient to effectively recruit RNF168 and mediate effective ubiquitin-conjugate production at DSB sites. Thus, we speculate that RNF8, RNF168 and/or BRCA1/BARD1 might require pre-SUMOylation of their targets and/or that SUMOylation regulates their ubiquitin-ligase activities. Future studies will be required to define the precise mechanisms by which the ubiquitin- and SUMO-conjugation systems cooperate at DSB sites, and determine how PIAS1 and PIAS4 impinge on chromatin structure, promote DSB signalling and repair, and potentially regulate yet other aspects of the DDR.

## METHODS SUMMARY

U2OS-based cell lines were maintained under standard conditions. cDNA cloning was by standard procedures. siRNA transfections were with Lipofectamine RNAiMAX (Invitrogen). Ionizing radiation was administered with a Faxitron X-ray machine (Faxitron X-ray Corporation). ATM inhibition was by KU-55933 (KuDOS Pharmaceuticals). Laser micro-irradiation was with a FluoView 1000 confocal microscope (Olympus) with 37 °C heating stage (Ibidi) and 405 nm diode (6 mW). FRAP was performed when laser-track accumulation of GFP-tagged protein reached maximal steady-state level. For immunofluorescence, cells were pre-extracted or not, fixed with 2% paraformaldehyde, permeabilized and stained. For whole cell extracts, cells were lysed on plates with 2% SDS, 50 mM Tris-HCl pH 7.5, 20 mM *N*-ethylmaleimide (Sigma-Aldrich) and protease inhibitor cocktail (Roche). To immunoprecipitate 53BP1, BRCA1 and SUMOylated proteins, different lysis and binding buffers were used (Supplementary Information). Homologous recombination and NHEJ assays were performed as described previously[17,28]. For ionizing radiation survival, cells were transfected with siRNA and exposed to ionizing radiation. After 10–14 days, colonies were stained with 0.5% crystal violet/20% ethanol, counted and normalized to plating efficiencies. For fluorescence-activated cell sorting of propidium iodide-stained cells, data were analysed by FlowJo software. All error bars represent s.d.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 24 March; accepted 4 November 2009.**


1. Stucki, M. & Jackson, S. P. MDC1/NFBD1: a key regulator of the DNA damage response in higher eukaryotes. *DNA Repair (Amst.)* **3,** 953–957 (2004).
2. Downs, J. A., Nussenzweig, M. C. & Nussenzweig, A. Chromatin dynamics and the preservation of genetic information. *Nature* **447,** 951–958 (2007).
3. Hay, R. T. SUMO: a history of modification. *Mol. Cell* **18,** 1–12 (2005).
4. Meulmeester, E. & Melchior, F. Cell biology: SUMO. *Nature* **452,** 709–711 (2008).
5. Bergink, S. & Jentsch, S. Principles of ubiquitin and SUMO modifications in DNA repair. *Nature* **458,** 461–467 (2009).
6. Geoffroy, M. C. & Hay, R. T. An additional role for SUMO in ubiquitin-mediated proteolysis. *Nature Rev. Mol. Cell Biol.* **10,** 564–568 (2009).
7. Doil, C. *et al.* RNF168 binds and amplifies ubiquitin conjugates on damaged chromosomes to allow accumulation of repair proteins. *Cell* **136,** 435–446 (2009).
8. Huen, M. S. *et al.* RNF8 transduces the DNA-damage signal via histone ubiquitylation and checkpoint protein assembly. *Cell* **131,** 901–914 (2007).
9. Mailand, N. *et al.* RNF8 ubiquitylates histones at DNA double-strand breaks and promotes assembly of repair proteins. *Cell* **131,** 887–900 (2007).
10. Morris, J. R. & Solomon, E. BRCA1: BARD1 induces the formation of conjugated ubiquitin structures, dependent on K6 of ubiquitin, in cells during DNA replication and repair. *Hum. Mol. Genet.* **13,** 807–817 (2004).
11. Stewart, G. S. *et al.* The RIDDLE syndrome protein mediates a ubiquitin-dependent signaling cascade at sites of DNA damage. *Cell* **136,** 420–434 (2009).
12. Evdokimov, E., Sharma, P., Lockett, S. J., Lualdi, M. & Kuehn, M. R. Loss of SUMO1 in mice affects RanGAP1 localization and formation of PML nuclear bodies, but is not lethal as it can be compensated by SUMO2 or SUMO3. *J. Cell Sci.* **121,** 4106–4113 (2008).
13. Lukas, C., Falck, J., Bartkova, J., Bartek, J. & Lukas, J. Distinct spatiotemporal dynamics of mammalian checkpoint regulators induced by DNA damage. *Nature Cell Biol.* **5,** 255–260 (2003).
14. Limoli, C. L. & Ward, J. F. A new method for introducing double-strand breaks into cellular DNA. *Radiat. Res.* **134,** 160–169 (1993).
15. Matsuoka, S. *et al.* ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316,** 1160–1166 (2007).
16. Hickson, I. *et al.* Identification and characterization of a novel and specific inhibitor of the ataxia-telangiectasia mutated kinase ATM. *Cancer Res.* **64,** 9152–9159 (2004).
17. Sartori, A. A. *et al.* Human CtIP promotes DNA end resection. *Nature* **450,** 509–514 (2007).
18. Potts, P. R. & Yu, H. Human MMS21/NSE2 is a SUMO ligase required for DNA repair. *Mol. Cell. Biol.* **25,** 7021–7032 (2005).
19. Kolas, N. K. *et al.* Orchestration of the DNA-damage response by the RNF8 ubiquitin ligase. *Science* **318,** 1637–1640 (2007).
20. Lou, Z. *et al.* MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. *Mol. Cell* **21,** 187–200 (2006).
21. Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. & Elledge, S. J. MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature* **421,** 961–966 (2003).
22. Wang, B. & Elledge, S. J. Ubc13/Rnf8 ubiquitin ligases control foci formation of the Rap80/Abraxas/Brca1/Brcc36 complex in response to DNA damage. *Proc. Natl Acad. Sci. USA* **104,** 20759–20763 (2007).
23. Xie, A. *et al.* Distinct roles of chromatin-associated proteins MDC1 and 53BP1 in mammalian double-strand break repair. *Mol. Cell* **28,** 1045–1057 (2007).
24. Aravind, L. & Koonin, E. V. SAP – a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem. Sci.* **25,** 112–114 (2000).
25. Durant, S. T. & Nickoloff, J. A. Good timing in the cell cycle for precise DNA repair by BRCA1. *Cell Cycle* **4,** 1216–1222 (2005).
26. Gudmundsdottir, K. & Ashworth, A. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* **25,** 5864–5874 (2006).
27. Pierce, A. J., Hu, P., Han, M., Ellis, N. & Jasin, M. Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. *Genes Dev.* **15,** 3237–3242 (2001).
28. Stucki, M. *et al.* MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell* **123,** 1213–1226 (2005).
29. Polanowska, J., Martin, J. S., Garcia-Muse, T., Petalcorin, M. I. & Boulton, S. J. A conserved pathway to activate BRCA1-dependent ubiquitylation at DNA damage sites. *EMBO J.* **25,** 2178–2188 (2006).
30. Tahk, S. *et al.* Control of specificity and magnitude of NF-κB and STAT1-mediated gene activation through PIASy and PIAS1 cooperation. *Proc. Natl Acad. Sci. USA* **104,** 11643–11648 (2007).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank S.P.J. lab members for support, in particular J. Harrigan, P. Huertas, S. Gravel, K. Dry and R. Chapman. We also thank C. Lukas for U2OS cells expressing GFP–BRCA1/Flag–BARD1, D. Durocher and G. Stewart for hTERT RIDDLE syndrome fibroblasts complemented with vector or HA–RNF168 and RNF168 antibody, T. Halazonetis for RNF8 antibody, R. Baer for the Flag–BARD1 construct, P. Harkin for the HA–BRCA1 construct, K. Iwabuchi for HA-tagged, full length, N, C, CΔBRCT and BRCT 53BP1 constructs, R. Walker for help with FACS, and J. R. Morris for sharing results before publication. Research in the S.P.J. lab is supported by grants from Cancer Research UK and the European Union (Integrated Project DNA repair, LSHG-CT-2005-512113, and Genomic Instability in Cancer and Precancer, HEALTH-F2-2007-201630).

**Author Contributions** R.B. cloned the PIAS cDNAs, tested the original siRNA efficiencies and provided help with processing of laser experiments. J.C. intensively helped with cell survival, homologous recombination and NHEJ experiments and provided support with tissue culture maintenance and stable-cell-line generation. S.P. set up the laser system in the laboratory and helped perform and analyse the FRAP experiments. K.M.M. provided the initial results on 53BP1 IRIF in PIAS4-depleted cells and constructed siRNA-resistant RFP–PIAS4. Y.G. initiated the project, led the teamwork and performed all other experiments described in the manuscript. Y.G. and S.P.J. conceived the study and wrote the paper. All authors discussed and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.P.J. (s.jackson@gurdon.cam.ac.uk).

## METHODS

**Cell culture.** U2OS cells were grown in DMEM (Sigma-Aldrich) supplemented with 10% fetal bovine serum (Biosera), 100 units/ml of penicillin, and 100 mg ml$^{-1}$ of streptomycin (Sigma-Aldrich). U2OS cells stably expressing GFP–MDC1 (ref. 1), GFP–53BP1, GFP–CtIP (ref. 2), GFP–RNF8, GFP–SUMO1, GFP–SUMO2, GFP–SUMO3, GFP–UBC9, GFP–PIAS1 (wild type, LD, ΔSAP and LDΔSAP) and RFP–PIAS4 (wild type, wild type siRNA resistant, LD, ΔSAP and LDΔSAP) were grown in standard U2OS medium supplemented with 1 mg ml$^{-1}$ of G418 (Gibco, Invitrogen). U2OS cells stably expressing GFP–BRCA1 and Flag–BARD1 (ref. 3) were provided by C. Lukas and were grown with 0.4 mg ml$^{-1}$ of G418. hTERT RIDDLE syndrome fibroblasts complemented with vector or HA–RIDDLIN/RNF168 (ref. 4) were provided by D. Durocher and G. Stewart, and were grown in standard medium supplemented with 0.5 mg ml$^{-1}$ of G418.

**siRNA transfection and sequences.** siRNA duplexes were obtained from MWG biotech or QIAGEN (Supplementary Table 1). Two consecutive rounds of siRNA transfections were carried out with Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's protocol unless otherwise specified. siRNA-transfected cells were assayed 48 h after transfection. For co-transfection with siRNA and expression constructs, cells were first transfected with siRNA followed by plasmid transfection 24 h later by using Fugene6 (Roche) according to manufacturer's protocol. Cells were assayed 48 h after plasmid transfections. siRNA-mediated downregulation of overexpressed protein was achieved by a first round of siRNA transfection as described above with an additional siRNA transfection 24 h after plasmid transfection. All PIAS4 siRNA mediated down regulation experiments were carried out using PIAS4-1 siRNA unless otherwise specified.

**Laser micro-irradiation and imaging of live and fixed cells.** For generation of localized damage in cellular DNA by exposure to an ultraviolet-A laser beam[5,6], cells were plated on glass-bottomed dishes (Willco-Wells) and pre-sensitized with 10 μM 5-bromo-2'-deoxyuridine (BrdU, Sigma-Aldrich) in phenol red-free medium (Invitrogen) for 24 h at 37 °C. Laser micro-irradiation was performed by using a FluoView 1000 confocal microscope (Olympus) equipped with a 37 °C heating stage (Ibidi) and a 405 nm laser diode (6 mW) focused through a ×60 UPlanSApo/1.35 oil objective to yield a spot size of 0.5–1 μm. The time of cell exposure to the laser beam was around 250 ms (fast scanning mode). Laser settings (0.40 mW output, 50 scans) were chosen that generate a detectable damage response restricted to the laser path in a pre-sensitization-dependent manner without noticeable cytotoxicity. Imaging of live and fixed cells was done on the same microscope by using the objective lens and software described above.

**Fluorescent recovery after photobleaching (FRAP).** FRAP analyses were performed on the microscope used for laser micro-irradiation when the accumulation of the GFP-tagged protein on the laser track reached its maximal steady-state level. After a series of three pre-bleach images, a rectangular region placed over the laser-damaged line was subjected to a bleach pulse (five scans with 488 nm argon laser focused through a ×60 UPlanSApo/1.35 oil objective, main scanner, 100% AOTF acousto-optical tunable filter, slow scanning mode), followed by image acquisition in 5 s intervals for GFP–53BP1 and at fastest speed for GFP–BRCA1. Average fluorescent intensities in the bleached region were normalized against intensities in an undamaged nucleus in the same field after background subtraction to correct for overall bleaching of the GFP signal due to repetitive imaging. For mathematical modelling of GFP-tagged protein mobility, $(I_t - I_0)/I_{pre}$ values were plotted as a function of time, where $I_0$ is the fluorescence intensity immediately after bleaching and $I_{pre}$ is the average of the three pre-bleach measurements. Estimation of mobile protein fraction ($A$) and residence time ($\tau$) were performed using Prism 4 software assuming the existence of one protein population using the following equation: $y(t) = A(1 - \exp(-t/\tau))$.

**Immunofluorescence.** Cells were washed three times with PBS 0.1% Tween-20 followed by pre-extraction for 10 min (pre-extraction buffer: 25 mM HEPES 7.4, 50 mM NaCl, 1 mM EDTA, 3 mM $MgCl_2$, 300 mM sucrose and 0.5% TritonX-100). Cells were fixed with 2% formaldehyde (w/v) in PBS for 20 min. Following three washes with PBS 0.1% Tween-20, cells were blocked for 1 h with 5% BSA in PBS 0.1% Tween-20 co-stained with the appropriate antibodies (Supplementary Table 3) in blocking solution over night and then co-immunostained with the appropriate secondary antibodies (Supplementary Table 3) in blocking solution. For imaging RNF168/RIDDLIN pre-extraction buffer: 20 mM HEPES, 20 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 0.5% NP-40, serine/threonine phosphatases inhibitor cocktail (Sigma-Aldrich), protease inhibitor cocktail (Roche). Pre-extraction step was omitted and permeabilization (0.5% Triton X-100 in PBS) was performed after fixation for imaging of UBC9, PIAS1, PIAS4, CtIP and RNF8.

**Treatment with small-molecule inhibitors and DNA-damaging agents.** Camptothecin was obtained from Sigma-Aldrich, ATM KU-55933 inhibitor was provided by KuDOS Pharmaceuticals. Ionizing radiation treatment was performed by using a Faxitron X-ray machine (Faxitron X-ray Corporation). Where appropriate, ATM inhibitor (20 μM) was applied to the culture medium 1 h before laser micro-irradiation.

**Plasmids and cloning.** SUMO1, SUMO2, SUMO3, UBC9, PIAS1, PIAS2, PIAS3 and PIAS4 were PCR amplified from a human fetal brain cDNA library and cloned into pCS2-mRFP (R. Y. Tsien) and pEGFP-C1 (Clontech). PIAS1ΔSAP and PIAS4ΔSAP were sub-cloned from the original clones whereas PIAS1C351A and PIAS4C342A/C347A were created using a QuikChange Site-Directed Mutagenesis kit (Stratagene). PIAS4 siRNA resistant clone was obtained by inserting the 7-nucleotide mismatches underlined (GATCCAAA GTCCGGACTGAA) into PIAS4 cDNA using a QuikChange Site-Directed Mutagenesis kit (Stratagene). MMS21 was cloned into pCL-NCX (Imgenex) initially modified to contain a 7His-3×Flag tag using an adaptor duplex. RNF8 was cloned into pCDNA3.1 (Invitrogen) initially modified to contain GFP. PIAS1 and PIAS4 were also cloned into pCDNA3.1(-) (Invitrogen) initially modified to contain 3×Flag-S-tag in the same reading frame as pEGFP-C1. Mammalian expression plasmids encoding HA–53BP1 (full length, N, C, CΔBRCT and BRCT) were provided by K. Iwabuchi. Mammalian expression plasmids encoding Flag–BARD1 and HA–BRCA1 were provided by R. Baer and P. Harkin respectively. Primers were obtained from Sigma-Aldrich (Supplementary Table 2).

**Immunoprecipitation and immunoblotting.** Cell extracts were prepared on plates by using lysis buffer containing 2% SDS, 50 mM Tris-HCl pH 7.5, 20 mM N-ethylmaleimide (Sigma-Aldrich) and protease inhibitor cocktail (Roche). Sonicating or passing the extracts 10 times through a 19G needle mounted syringe reduced viscosity. For 53BP1 immunoprecipitation, cell extracts were prepared as mentioned above and then diluted 1:20 with lysis buffer containing 150 mM NaCl and 1% NP40 instead of SDS. For BRCA1 immunoprecipitation, cell extracts were prepared by using lysis buffer containing 20 mM HEPES pH 7.4, 450 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 1% Tween20, 10% glycerol, serine/threonine phosphatase inhibitor cocktail (Sigma-Aldrich), protease inhibitor cocktail (Roche) and 10 mM N-ethylmaleimide (Sigma-Aldrich), the extracts were then sonicated and diluted 1:2 with the same buffer lacking NaCl. For GFP and HA–53BP1 immunoprecipitation using GFP-Trap-A (ChromoTek) and anti haemagglutinin antibodies, cell extracts were prepared by using lysis buffer containing 20 mM HEPES pH 7.4, 500 mM NaCl, 1.5 mM $MgCl_2$, 1 mM EGTA, 1% Triton X-100, 10% glycerol, serine/threonine phosphatase inhibitor cocktail (Sigma-Aldrich), protease inhibitor cocktail (Roche) and 10 mM N-ethylmaleimide (Sigma-Aldrich), the extracts were then sonicated and diluted 1:2 with the same buffer lacking NaCl. In all cases the extracts were cleared using centrifugation at 16,000 g for 45 min at 4 °C. Antibodies against 53BP1 (Sigma-Aldrich) and BRCA1 (Santa-Cruz 1:1 mix of rabbit polyclonals, same mix was used for immunoblotting) were pre-bound to Dynabeads ProteinG (Invitrogen) and incubated for 2 h at room temperature (53BP1) or overnight at 4 °C (BRCA1, GFP-Trap-A and HA) followed by five washes with immunoprecipitation buffer (two washes with lysis buffer were added for the GFP-Trap-A immunoprecipitation) and 5 min boiling in 1.5× SDS sample buffer. Proteins were resolved by 4–18% gradient SDS–PAGE (unless otherwise specified) and transferred to PVDF membrane (GE Healthcare). Immunoblotting was performed with the appropriate antibodies (Supplementary Table 3).

**Random plasmid integration assay.** Assays were performed as previously described[7] with minor modifications. Briefly, one day after transfection with siRNA, U2OS cells were transfected with BamHI–XhoI linearized pEGFP-C1 (Clontech). The following day, cells were collected, counted and plated on three plates, one of which contained 0.5 mg ml$^{-1}$ G418. One day after plating, the cells on a plate lacking G418 were fixed to assess transfection efficiency and the other two plates were incubated for 10–14 days at 37 °C to allow colony formation. Colonies were stained with 0.5% crystal violet/20% ethanol and counted. Random-plasmid integration events were normalized to transfection and plating efficiencies.

**Homologous recombination assay.** A U2OS clone with the integrated homologous recombination reporter DR-GFP was generated as described previously[2,8]. One day after transfection with siRNA, U2OS-DR-GFP cells were co-transfected with an I-SceI expression vector (pCBA-I-SceI) and a vector expressing monomeric red fluorescent protein (pCS2-mRFP). The latter plasmid was added in a 1:5 ratio to mark the I-SceI-positive cells. Cells were harvested one day after I-SceI transfection and subjected to flow cytometric analysis to examine recombination induced by I-SceI digestion. Only RFP-positive cells were analysed for homologous recombination efficiency to circumvent possible differences in transfection efficiencies. Fluorescence-activated cell sorting data were analysed by using Summit V4.3 software to reveal the percentage of GFP-positive cells relative to the number of transfected cells (RFP positive). The data were related to a control siRNA treatment in each individual experiment. The dividing line

between GFP (homologous recombination) positive and negative cells was set to 0.5% background level of GFP-positive cells in the internal control (RFP positive, not transfected with I-SceI). This gate was then applied to the RFP/I-SceI positive samples to determine homologous recombination efficiency. Results are presented as a percentage of control siRNA.

**Ionizing radiation survival assays.** U2OS cells were transfected with siRNA and exposed to ionizing radiation. Cells were left for 10–14 days at 37 °C to allow colony formation. Colonies were stained with 0.5% crystal violet/20% ethanol and counted. Results were normalized to plating efficiencies.

**Florescence-activated cell sorting (FACS).** To determine cell-cycle distribution, cells were fixed with 70% ethanol, incubated for 30 min with RNase A (250 $\mu g\,ml^{-1}$) and propidium iodide (10 $\mu g\,ml^{-1}$) at 37 °C and analysed by FACS. Data were analysed by using FlowJo software to reveal the percentage of cells in each phase of the cell cycle.

LETTERS

# Coordinating DNA replication by means of priming loop and differential synthesis rate

Manjula Pandey[1], Salman Syed[2], Ilker Donmez[1], Gayatri Patel[1], Taekjip Ha[2,3] & Smita S. Patel[1]

**Genomic DNA is replicated by two DNA polymerase molecules, one of which works in close association with the helicase to copy the leading-strand template in a continuous manner while the second copies the already unwound lagging-strand template in a discontinuous manner through the synthesis of Okazaki fragments[1,2]. Considering that the lagging-strand polymerase has to recycle after the completion of every Okazaki fragment through the slow steps of primer synthesis and hand-off to the polymerase[3–5], it is not understood how the two strands are synthesized with the same net rate[6–9]. Here we show, using the T7 replication proteins[10,11], that RNA primers are made 'on the fly' during ongoing DNA synthesis and that the leading-strand T7 replisome does not pause during primer synthesis, contrary to previous reports[12,13]. Instead, the leading-strand polymerase remains limited by the speed of the helicase[14]; it therefore synthesizes DNA more slowly than the lagging-strand polymerase. We show that the primase–helicase T7 gp4 maintains contact with the priming sequence during ongoing DNA synthesis; the nascent lagging-strand template therefore organizes into a priming loop that keeps the primer in physical proximity to the replication complex. Our findings provide three synergistic mechanisms of coordination: first, primers are made concomitantly with DNA synthesis; second, the priming loop ensures efficient primer use and hand-off to the polymerase; and third, the lagging-strand polymerase copies DNA faster, which allows it to keep up with leading-strand DNA synthesis overall.**

To investigate the functional cooperativity between the enzymatic activities of the T7 replication complex, we measured the kinetics of DNA unwinding, DNA synthesis and primer synthesis on synthetic replication-fork substrates with and without the T7 priming sequence (3′-CTGGG-5′; Supplementary Table 1). Efficient synthesis of RNA primers from dimer to pentamer by T7 replisome (T7 gp4 and T7 DNA polymerase) was observed on the priming fork (Fig. 1a and Supplementary Fig. 1) with a half-life of about 0.5 s and a yield of more than 60% (Fig. 1a, right). T7 gp4 alone also makes RNA primers on this priming fork, but roughly tenfold more slowly (Supplementary Fig. 1), which is consistent with polymerase assistance of the helicase rate[14]. An average 46% yield of primer synthesis with forks of different lengths and sequences (Supplementary Table 2) indicates that T7 replisome lays down primers on newly unwound lagging-strand template with a high efficiency. In addition, the newly made primers are elongated through lagging-strand DNA synthesis (Supplementary Fig. 1).

All-or-none DNA strand separation assays[15,16] under primer synthesis conditions show that T7 replisome unwinds the priming fork and the control fork (without the priming sequence) with similar rate constants at all dTTP concentrations (Supplementary Fig. 2). Single-molecule fluorescence resonance energy transfer (FRET) unwinding assays[17] show an increase in fluorescence intensity of Cy3 (donor, green) and a decrease in Cy5 (acceptor, red) fluorescence intensity as a result of DNA unwinding and synthesis (Fig. 1b), and the priming and control forks show a similar FRET decrease time ($\Delta t$) of $0.4 \pm 0.027$ and $0.37 \pm 0.022$ s, respectively. By comparing the histograms of FRET values visited during the unwinding reactions between the T7 replisome reaction and the roughly threefold slower reaction by T7 gp4 alone (Supplementary Fig. 3), we further confirmed that the T7 replisome does not pause during primer synthesis (Supplementary Fig. 6). If the T7 replisome paused for several seconds every time a primer is made[12], our single-molecule analysis would have detected the pausing events.

To investigate whether DNA synthesis was occurring concomitantly with primer synthesis, the kinetics of strand-displacement DNA synthesis was measured on the priming and control forks under the same reaction conditions as in Fig. 1a. In the high-resolution sequencing gels used to analyse the DNA synthesis kinetics, any pausing of the T7 replisome activity as a result of primer synthesis would be detected as an accumulation of specific DNA products in the priming-fork reactions, but not in the control. However, no unusual accumulation of intermediate DNA products as a result of replisome pausing was observed with the priming-fork template in comparison with the control (Fig. 1c). Consistent with this result, the average DNA elongation rates on priming ($126 \pm 9$ nucleotides $s^{-1}$) and control ($113 \pm 8$ nucleotides $s^{-1}$) forks were similar (Fig. 1c). No pausing was detected on longer forks (Supplementary Figs 1 and 8) or on forks with different GC contents or at different dNTP concentrations (Supplementary Table 3). Finally, coupled leading-strand and lagging-strand DNA synthesis measured by the rolling-circle assay with T7 gp2.5 showed no effect of lagging-strand synthesis on the rate of leading-strand synthesis (Supplementary Fig. 9). Similar observations have been made with the T4 replication proteins[18], although a recent study of *Escherichia coli* replication yielded a different result[19]. Overall, our results indicate that DNA synthesis continues uninterrupted while RNA primers are laid down, and the leading-strand polymerase does not slow as a function of primase activity or as a result of any of the steps during lagging-strand polymerase recycling.

Because DNA synthesis continues uninterrupted while primers are being synthesized, our results predict that the nascent lagging-strand template should loop out between the covalently linked helicase and primase domains of T7 gp4 (Fig. 2A)[20]. The formation of such a priming loop during DNA synthesis was probed with single-molecule FRET experiments: Cy3 and Cy5 fluorophores were introduced 40 base pairs (bp) apart on the lagging-strand template of the surface-attached DNA fork (Fig. 2B). Before the unwinding of DNA, no FRET was observed because of the long (40-bp) distance between the fluorophores (Fig. 2B, C, a). As T7 replisome unwinds the double-stranded DNA (dsDNA), the donor Cy3 shows an increase in intensity (green trace in Fig. 2C, top panel) due to a change in environment from

[1]Department of Biochemistry, University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School, Piscataway, New Jersey 08854, USA. [2]Howard Hughes Medical Institute, Urbana, Illinois 61801, USA. [3]Department of Physics and the Center for the Physics of Living Cells, University of Illinois, Urbana-Champaign, Illinois 61801, USA.

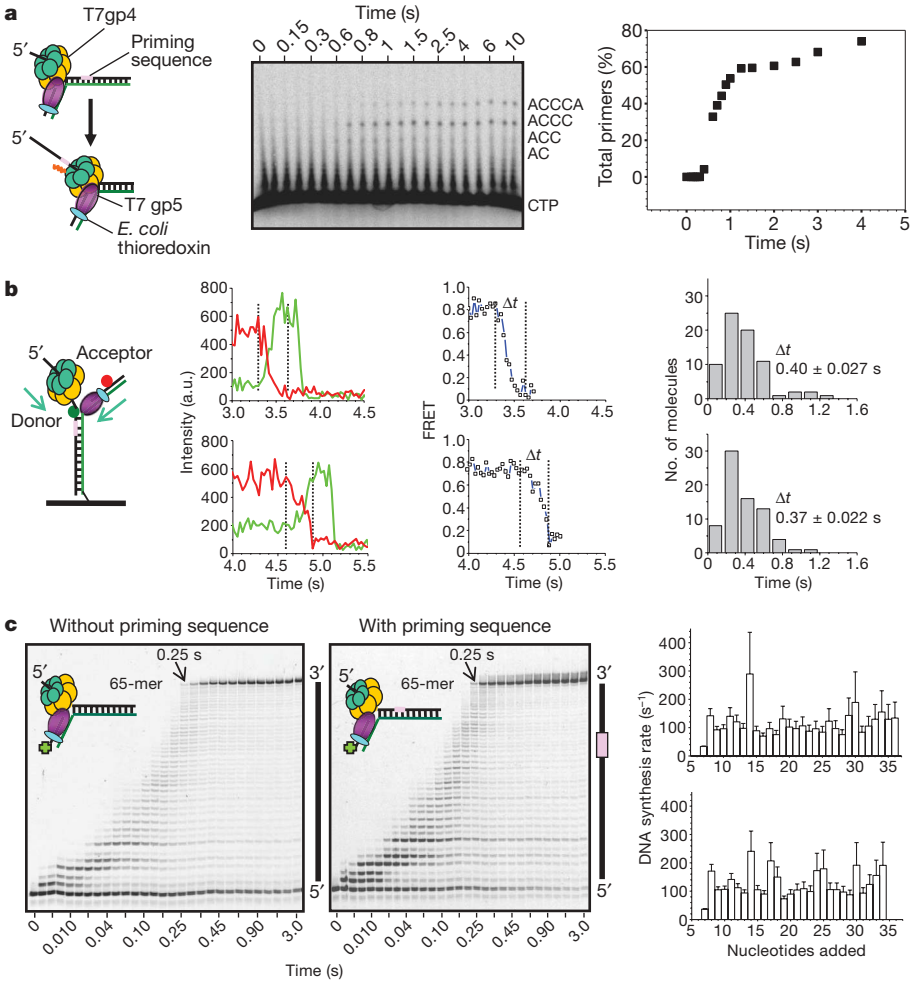



**Figure 1 | Primer synthesis occurs concomitantly with DNA unwinding and synthesis.** **a**, Middle: sequencing gel showing the time course of RNA primers made by T7 replisome (600 nM) on priming ts40P-50% fork DNA substrate (300 nM) with dNTPs (1 mM each), CTP (0.5 mM), [α-$^{32}$P]CTP, ATP (1 mM), $MgCl_2$ (4 mM free) and $dT_{90}$ trap (3 μM) at 18 °C in replication buffer. Right: plot of total yield of RNA primers against time. **b**, Left graphs: representative Cy3 and Cy5 intensity traces showing DNA unwinding on priming (top) and control without the priming sequence (bottom) substrates by single-molecule FRET with T7 replisome (50 nM), dNTPs (1 mM each), 1 mM ATP, 1 mM CTP and 4 mM free $Mg^{2+}$, at 23 ± 1 °C. Additional traces are shown in Supplementary Fig. 4. Middle graphs: FRET time courses; right graphs: dwell-time histograms; determinations are described in Methods. Priming and control substrates show similar FRET decrease times ($\Delta t$) due to unwinding and synthesis. **c**, Polyacrylamide sequencing gels show progressive elongation of fluorescein-labelled 24-mer DNA primer by T7 replisome on control ts40-50% (left) and priming ts40P-50% DNA fork substrates (middle) under the same conditions as in **a**. Nucleotide incorporation rates (right) were calculated from the global fit to the polymerization model (Supplementary Information); the average DNA primer elongation rates were 113 ± 8 and 126 ± 9 nucleotides $s^{-1}$ (means ± s.e.m.) in the absence (top) and presence (bottom) of priming sequence, respectively. Error bars indicate s.e.m.

protein-induced fluorescence enhancement[21] and DNA strand separation[22] (Fig. 2B, b). Continued DNA unwinding brings the priming sequence and the donor nearby close to the primase domain, where they are held in place. The replisome continues unwinding the DNA while the primase domain is engaged with the priming sequence; therefore, at some point in time, the acceptor comes close to the donor (Fig. 2B, c) and this was detected as an increase in FRET (Fig. 2C, c, top and middle panels), providing evidence for the formation of a priming loop. In all, 40 molecules out of about 75 with a fluorescently active donor and acceptor showed formation of a priming loop. With continued unwinding, the priming loop grows in size and the donor and acceptor move apart (Fig. 2B, d), which was detected as a decrease in FRET (Fig. 2C, d, top and middle panels). Finally, the total fluorescence signal disappears on completion of the reaction and the release of the fluorescently labelled DNA strand from the surface.

The control fork showed an increase in donor intensity (green) (Fig. 2C, bottom panel) but no increase in acceptor intensity or FRET (more than 200 molecules were analysed). The donor intensity time (Fig. 2D, bottom panel)—the time between the jump in donor intensity and the total signal disappearance—was the same for priming and control DNAs, indicating that both DNAs are unwound at the same rate. The average time for FRET increase (Fig. 2D, top panel) with the priming fork was threefold longer for the experiments with T7 gp4 only (Supplementary Fig. 11) compared with T7 replisome, indicating the assistance of the polymerase in the reactions[14]. To test whether compaction of unwound single-stranded DNA (ssDNA) might give rise to high FRET values[23], we measured FRET between Cy3 and Cy5 separated by 40 nucleotides of ssDNA and found the average FRET value to be 0.2, much lower than those obtained during priming-loop formation (Supplementary Fig. 12). Overall, these results show, first, that the primase remains engaged with the priming sequence while DNA continues to be synthesized, and second, that the nascent lagging strand forms a priming loop.

Making primers ahead of time during ongoing DNA synthesis minimizes the delay due to primer synthesis, and keeping the RNA primers in physical proximity to the replicating complex provides a mechanism for efficient use of the priming sequence and hand-off to the polymerase. Nevertheless, lagging-strand polymerase dissociation, primer hand-off and initiation of a new Okazaki fragment synthesis event take time that could delay the synthesis of the lagging strand. Because the leading-strand replisome does not slow or pause during primer synthesis, the question remains as to how the lagging-strand polymerase keeps up with the leading-strand polymerase after many Okazaki fragment synthesis events. We therefore tested an alternative model, proposed more than 20 years ago[24], that the leading-strand polymerase simply moves with a slower overall rate than the lagging-strand polymerase.

The transient-state kinetic assays allow us to measure the DNA synthesis rates very precisely. To measure the rate of DNA synthesis as catalysed by the lagging-strand polymerase, we used a primer–template DNA substrate coated with T7 gp2.5 that mimics the already unwound lagging-strand template. To measure the rate of leading-strand synthesis by the T7 replisome, we used a fork substrate that contained the same template sequence in the dsDNA. T7 DNA polymerase copied the gp2.5-coated ssDNA template 30% faster (188 ± 10 nucleotides $s^{-1}$) than the T7 replisome did (132 ± 10 nucleotides $s^{-1}$) (Fig. 3). A faster rate of DNA synthesis by T7 DNA polymerase than that of the replisome was observed with *E. coli* single-strand binding protein-coated template (158 ± 10 nucleotides $s^{-1}$) and without any single-strand binding protein (200 ± 8 nucleotides $s^{-1}$) (Supplementary Fig. 13 and Supplementary Table 3). That the T7 replisome moves more slowly than the DNA polymerase alone is consistent with the

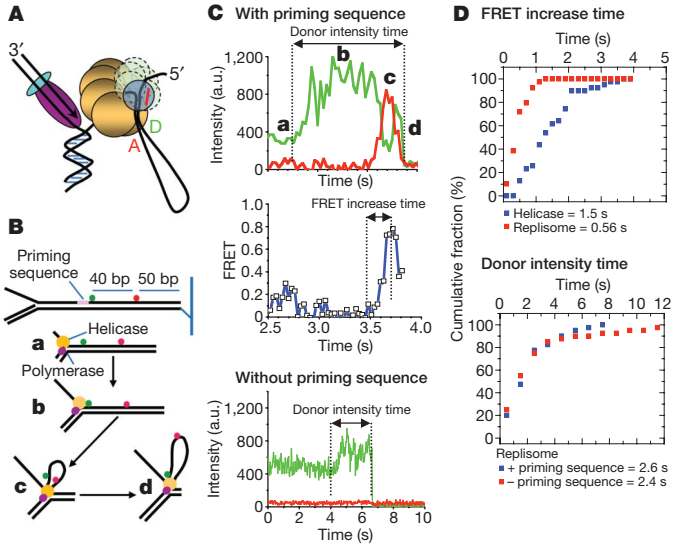



**Figure 2 | Priming loop: the primase domain maintains contact with the priming sequence during replication.** **A**, Diagram showing that while the primase domain remains bound to the priming sequence to make RNA (red), nascent lagging-strand template forms a priming loop. Red A, acceptor; green D, donor. **B**, Fluorescently labelled DNA fork to investigate the priming loop. **C**, Top: Cy3 (donor, green) and Cy5 (acceptor, red) intensity time traces during DNA synthesis by T7 replisome. Labels **a**–**d** correspond to the DNA states in the diagrams in **B**. Middle: plot of FRET efficiency against time. Bottom: representative Cy3 and Cy5 intensity traces on the control fork without the priming sequence. **D**, Plots of cumulative fraction against the indicated time intervals determined from single-molecule time traces. The time intervals measured are marked by the arrows in **C**. In all, 40 molecules from one experiment were used to build the curves (see Methods). Supplementary Fig. 10 shows additional representative time traces.

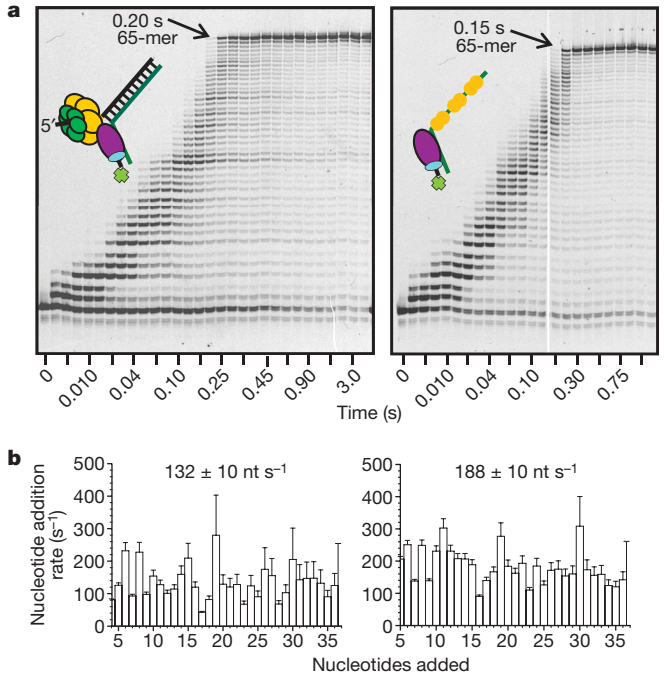



**Figure 3 | Lagging-strand synthesis is faster than leading-strand synthesis.** **a**, Sequencing gels showing the progressive elongation of fluorescein-labelled DNA primer by T7 replisome (200 nM) on 100 nM of fork DNA substrate ts40 (left) and primer-template DNA substrate p/t40 (right) with T7 gp2.5 (5 μM) under the same conditions as in Fig. 1a, without $dT_{90}$ trap. **b**, The rate constants of individual nucleotide (nt) addition steps. Nucleotide incorporation rates were calculated from the global fits to the polymerization model; the average DNA primer elongation rates were 132 ± 10 (left) and 188 ± 10 nucleotides $s^{-1}$ (means ± s.e.m.). Error bars indicate s.e.m.

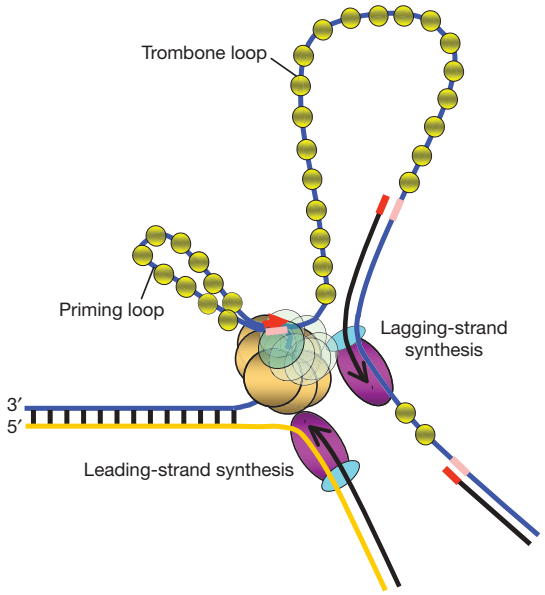



**Figure 4 | Model of T7 DNA replication.** The leading-strand template (yellow) is copied continuously by the cooperative action of T7 DNA polymerase and T7 gp4 while the lagging-strand template (in blue, coated with gp2.5) is copied by T7 DNA polymerase through the synthesis of Okazaki fragments. The physical coupling of the leading-strand and lagging-strand polymerases by interactions between T7 gp4 and gp2.5 creates a trombone loop. The priming loop (coated with gp2.5) is created between the physically linked primase and helicase domains of T7 gp4 as a result of ongoing DNA synthesis during primer synthesis. The priming loop keeps the nascent primer within physical reach of the lagging-strand polymerase. On primer hand-off to the lagging-strand polymerase, the priming loop becomes part of the trombone loop.

notion that the DNA synthesis rate is limited by the speed of the helicase[14]. From multiple experiments we estimate that T7 DNA polymerase alone copies the ssDNA template on average 38% faster than the T7 replisome does (Supplementary Table 3). From the 30% difference in rates, we calculate that the leading-strand polymerase will take on average 6–7 s longer than the lagging-strand polymerase to copy 3,000 bp of DNA, the average length of an Okazaki fragment. Unless physical coupling slows its rate, the lagging-strand polymerase will reach the end of the previously made Okazaki fragment with 6–7 s to spare to pick up a new primer and initiate another round of Okazaki fragment synthesis.

Our not observing replisome pausing is in contrast with previous reports[12,13] suggesting that T7 replisome pauses during primer synthesis. DNA synthesis in those studies was measured indirectly by following the overall shortening of DNA as dsDNA was converted to coiled ssDNA. In the presence of ATP plus CTP, intervals of 5–6 s were observed with no change in DNA length, which was attributed to replisome stopping. We propose that these pauses are caused not by replisome stopping but by the conversion of ssDNA back to dsDNA as a result of uncoupled lagging-strand synthesis. Although reactions were washed, contaminating polymerase catalysing uncoupled DNA synthesis including those tethered by means of T7 gp4 (ref. 25) could not be ruled out. Under conditions in which excess polymerase was present, both transient loops and pauses were observed[12]. The possibility could not be ruled out that the pausing and looping pattern was caused by separate Okazaki fragment synthesis events. Those that were coupled showed loop release, and those that were uncoupled showed the pausing behaviour.

On the basis of our studies, we propose that T7 replisome does not pause during primer synthesis or any of the steps of lagging-strand synthesis. Instead, the following synergistic mechanisms exist to coordinate strand synthesis. First, primers are made ahead of time during ongoing DNA synthesis; hence, primer synthesis itself does

not delay lagging-strand synthesis. Second, the primer is kept in physical proximity to the replication complex by means of a priming loop that ensures efficient primer use and hand-off to the lagging-strand polymerase (Fig. 4). Third, the lagging-strand polymerase copies the ssDNA template at a faster rate[24], providing extra time for the recycling steps. In addition to moving faster, multiple lagging-strand polymerases could work at the same time to complete lagging-strand synthesis in a shorter time[26]. Under certain conditions, the lagging-strand polymerase may jump to a new primer before completion of the Okazaki fragment, thereby leaving gaps that can be filled in later[27]. Because the basic mechanism of dsDNA replication is conserved from phage to humans[1,2], the mechanisms revealed from studies of the T7 replication proteins are broadly applicable to the more complex replication complexes of bacteria and eukaryotes.

## METHODS SUMMARY

**Ensemble kinetic assays.** T7 gp4 (ref. 28) and T7 DNA polymerase[29] (T7 gp5/*E. coli* thioredoxin) were preassembled on the DNA with dTTP and EDTA in replication buffer (50 mM Tris-HCl, pH 7.6, 40 mM NaCl, 10% glycerol) and the reactions were initiated with the addition of $MgCl_2$ and the rest of the dNTPs, with or without ATP, CTP, and $dT_{90}$ (90-nucleotide poly(dT) added to trap free and dissociated proteins). The kinetics of primer synthesis and DNA synthesis was measured with a rapid quenched-flow instrument (KinTek Corp.) and products were resolved on 24% or 25% polyacrylamide/urea sequencing gel. DNA synthesis kinetics were fitted to the polymerization model (Supplementary Information).

**Single-molecule FRET assays.** Single-molecule FRET experiments to measure unwinding and priming-loop formation were performed on a wide-field total-internal-reflection fluorescence microscope with 30 ms time resolution and imaged by means of a charge-coupled-device camera (iXon DV 887-BI; Andor Technology)[30]. The Cy3 and Cy5 fluorophores were internally labelled on the dT through a $C_6$ amino linker. Gel-based DNA synthesis reactions were performed to confirm that the fluorophores on the DNA did not affect DNA synthesis (Supplementary Fig. 7). The priming-loop substrates were prepared by ligating donor and acceptor labelled DNAs (Supplementary Information). FRET was calculated as the ratio of the acceptor intensity and the total (acceptor plus donor) intensity after correcting for cross-talk between the donor and acceptor channels and subtracting the background. For Figs 1b and 2, the initiation of FRET change and its saturation were scored by visual inspection of the donor and acceptor intensities and the time difference between the two points was designated as $\Delta t$. The calculated FRET efficiency from this method was demonstrated to be robust (Supplementary Fig. 5). The time for photobleaching of the fluorophores was at least tenfold longer than the unwinding time, and no unwinding-like signal was observed without the addition of $Mg^{2+}$.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 23 July; accepted 26 October 2009.**
**Published online 18 November 2009.**

1. Benkovic, S. J., Valentine, A. M. & Salinas, F. Replisome-mediated DNA replication. *Annu. Rev. Biochem.* **70,** 181–208 (2001).
2. O'Donnell, M. Replisome architecture and dynamics in *Escherichia coli*. *J. Biol. Chem.* **281,** 10653–10656 (2006).
3. Stukenberg, P. T., Turner, J. & O'Donnell, M. An explanation for lagging strand replication: polymerase hopping among DNA sliding clamps. *Cell* **78,** 877–887 (1994).
4. Frick, D. N. & Richardson, C. C. DNA primases. *Annu. Rev. Biochem.* **70,** 39–80 (2001).
5. Patel, S. S., Hingorani, M. M. & Ng, W. M. The K318A mutant of bacteriophage T7 DNA primase–helicase protein is deficient in helicase but not primase activity and inhibits primase–helicase protein wild-type activities by heterooligomer formation. *Biochemistry* **33,** 7857–7868 (1994).
6. Alberts, B. M. *et al.* Studies on DNA replication in the bacteriophage T4 *in vitro* system. *Cold Spring Harb. Symp. Quant. Biol.* **47,** 655–668 (1983).
7. Wu, C. A., Zechner, E. L. & Marians, K. J. Coordinated leading- and lagging-strand synthesis at the *Escherichia coli* DNA replication fork. I. Multiple effectors act to modulate Okazaki fragment size. *J. Biol. Chem.* **267,** 4030–4044 (1992).
8. Lee, J., Chastain, P. D. II, Kusakabe, T., Griffith, J. D. & Richardson, C. C. Coordinated leading and lagging strand DNA synthesis on a minicircular template. *Mol. Cell* **1,** 1001–1010 (1998).
9. Salinas, F. & Benkovic, S. J. Characterization of bacteriophage T4-coordinated leading- and lagging-strand synthesis on a minicircle substrate. *Proc. Natl Acad. Sci. USA* **97,** 7196–7201 (2000).
10. Donmez, I. & Patel, S. S. Mechanisms of a ring shaped helicase. *Nucleic Acids Res.* **34,** 4216–4224 (2006).
11. Hamdan, S. M. & Richardson, C. C. Motors, switches, and contacts in the replisome. *Annu. Rev. Biochem.* **78,** 205–243 (2009).
12. Lee, J. B. *et al.* DNA primase acts as a molecular brake in DNA replication. *Nature* **439,** 621–624 (2006).
13. Hamdan, S. M., Loparo, J. J., Takahashi, M., Richardson, C. C. & van Oijen, A. M. Dynamics of DNA replication loops reveal temporal control of lagging-strand synthesis. *Nature* **457,** 336–339 (2009).
14. Stano, N. M. *et al.* DNA synthesis provides the driving force to accelerate DNA unwinding by a helicase. *Nature* **435,** 370–373 (2005).
15. Donmez, I. & Patel, S. S. Coupling of DNA unwinding to nucleotide hydrolysis in a ring-shaped helicase. *EMBO J.* **27,** 1718–1726 (2008).
16. Jeong, Y. J., Levin, M. K. & Patel, S. S. The DNA-unwinding mechanism of the ring helicase of bacteriophage T7. *Proc. Natl Acad. Sci. USA* **101,** 7264–7269 (2004).
17. Ha, T. *et al.* Initiation and re-initiation of DNA unwinding by the *Escherichia coli* Rep helicase. *Nature* **419,** 638–641 (2002).
18. Yang, J., Trakselis, M. A., Roccasecca, R. M. & Benkovic, S. J. The application of a minicircle substrate in the study of the coordinated T4 DNA replication. *J. Biol. Chem.* **278,** 49828–49838 (2003).
19. Yao, N. Y., Georgescu, R. E., Finkelstein, J. & O'Donnell, M. E. Single-molecule analysis reveals that the lagging strand increases replisome processivity but slows replication fork progression. *Proc. Natl Acad. Sci. USA* **106,** 13236–13241 (2009).
20. Corn, J. E., Pelton, J. G. & Berger, J. M. Identification of a DNA primase template tracking site redefines the geometry of primer synthesis. *Nature Struct. Mol. Biol.* **15,** 163–169 (2008).
21. Myong, S. *et al.* Cytosolic viral sensor RIG-I is a 5′-triphosphate-dependent translocase on double-stranded RNA. *Science* **323,** 1070–1074 (2009).
22. Sanborn, M. E., Connolly, B. K., Gurunathan, K. & Levitus, M. Fluorescence properties and photophysics of the sulfoindocyanine Cy3 linked covalently to DNA. *J. Phys. Chem. B* **111,** 11064–11074 (2007).
23. Liu, S., Abbondanzieri, E. A., Rausch, J. W., Le Grice, S. F. & Zhuang, X. Slide into action: dynamic shuttling of HIV reverse transcriptase on nucleic acid substrates. *Science* **322,** 1092–1097 (2008).
24. Selick, H. E. *et al.* in *DNA Replication and Recombination* (eds McMacken, R. & Kelly, T. J.) 183–214 (Alan R. Liss Inc., 1987).
25. Hamdan, S. M. *et al.* A unique loop in T7 DNA polymerase mediates the binding of helicase-primase, DNA binding protein, and processivity factor. *Proc. Natl Acad. Sci. USA* **102,** 5096–5101 (2005).
26. McInerney, P., Johnson, A., Katz, F. & O'Donnell, M. Characterization of a triple DNA polymerase replisome. *Mol. Cell* **27,** 527–538 (2007).
27. Yang, J., Nelson, S. W. & Benkovic, S. J. The control mechanism for lagging strand polymerase recycling during bacteriophage T4 DNA replication. *Mol. Cell* **21,** 153–164 (2006).
28. Patel, S. S., Rosenberg, A. H., Studier, F. W. & Johnson, K. A. Large scale purification and biochemical characterization of T7 primase/helicase proteins. Evidence for homodimer and heterodimer formation. *J. Biol. Chem.* **267,** 15013–15021 (1992).
29. Patel, S. S., Wong, I. & Johnson, K. A. Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry* **30,** 511–525 (1991).
30. Ha, T. Single-molecule fluorescence resonance energy transfer. *Methods* **25,** 78–86 (2001).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank C. M. Drain for a critical reading of the paper, K. Picha and S. Patel for preparation of the minicircle DNA, and S. Arslan and K. S. Lee for help with single-molecule FRET data analysis. This work was supported by NIH grants GM55310 (S.S.P.) and GM065367 (T.H.) and NSF grants 0822613 and 0646550 (T.H.). T.H. is an investigator with the Howard Hughes Medical Institute.

**Author Contributions** M.P. purified T7 gp5 and T7 gp4, and constructed DNA substrates for the priming-loop studies, and obtained and analysed all the ensemble DNA synthesis and primer synthesis experiments. S.S. developed robust single-molecule assays for observing DNA unwinding and priming-loop formation, and obtained and analysed all single-molecule data. I.D. and G.P. performed the ensemble unwinding experiments. M.P., S.S., S.S.P. and T.H. designed the experiments, analysed the data and wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.H. (tjha@illinois.edu) or S.S.P. (patelss@umdnj.edu).

## METHODS

**Proteins and DNA.** T7 gp4A′ and gp5 (exo-) proteins were purified as described previously[28,29]. Thioredoxin from *E. coli* was purchased from Sigma-Aldrich. Protein concentration was calculated by ultraviolet absorption (in 8 M guanidinium chloride) using extinction coefficients at 280 nm of 0.0836 $\mu M^{-1} cm^{-1}$ for T7 gp4A′ and 0.13442 $\mu M^{-1} cm^{-1}$ for T7 gp5 (exo-). Oligodeoxynucleotides (Supplementary Table 1) were purchased from Integrated DNA Technology or Sigma-Aldrich, and purified by PAGE before use. Substrates for the stopped-flow and gel-based DNA-unwinding assays had an amino linker at the 5′ end of their bottom strands, which was labelled with 5-(and-6)-carboxyfluorescein succinimidyl ester, using the procedure from Molecular Probes with carbonate buffer. Proteins were preassembled on the DNA before the start of the reaction: T7 gp4 was added to the fork substrate with dTTP and EDTA in the replication buffer and incubated on ice for 30 min. For assembling T7 replisome, T7 DNA polymerase (T7 gp5 and *E. coli* thioredoxin (1:5) mixed for 5 min at 22 °C in replication buffer containing freshly made 5 mM dithiothreitol (DTT)[29]) was added to T7 gp4 and the DNA mixture and incubated at room temperature (23 ± 1 °C) for a further 30 min.

**RNA primer synthesis assay.** The protein–DNA complex was loaded in one syringe of the rapid quenched-flow instrument. The second syringe contained $MgCl_2$, ATP, CTP mixed with a trace amount of [α-$^{32}$P]CTP, dATP, dCTP, dGTP, and $dT_{90}$ trap in replication buffer (50 mM Tris-HCl, pH 7.6, 40 mM NaCl, 10% glycerol). The reaction was initiated by mixing equal volumes of the two solutions at 18 °C and quenched after various time intervals with 300 mM EDTA. Primers generated in the reaction were resolved on 25% polyacrylamide/3 M urea sequencing gel with 1.5 × TBE buffer, running the gel only three-quarters of the sequencing gel length. The gel was imaged on a Typhoon Phosphorimager, and the products were analysed with ImageQuant 5.0 software. The yield of RNA primer synthesis was determined from the radiolabelled CTP incorporation in the dimer to pentamer RNA, taking into account the number of C bases in the primers.

**DNA synthesis kinetics.** T7 replisome-fork DNA or T7 DNA polymerase–primer/template DNA complex was loaded in one syringe of the quenched-flow instrument. The second syringe contained dATP, dCTP and dGTP, with or without NTPs (ATP and CTP), $MgCl_2$ and trap (where applicable; see Supplementary Table 3) in replication buffer. Reactions were initiated by rapidly mixing equal volumes of the two solutions, and quenched after various intervals with 300 mM EDTA. The quenched solution was loaded on 22% or 24% polyacrylamide/7 M urea sequencing gel with 1.5 × TBE buffer. The gel was imaged with a Phosphorimager, and each DNA band was quantified with ImageQuant software. The time courses of the individual DNA product formations and decays were fitted to the polymerization model with *gfit* (http://gfit.sourceforge.net) to obtain the individual nucleotide addition rate constants from which the average DNA primer elongation rate was calculated.

**Single-molecule FRET experiments and data analysis.** Biotin was attached at the 5′ end of the DNA strand during DNA synthesis. Cy3 *N*-hydroxysuccinimido (NHS) ester and Cy5 NHS ester (GE Healthcare) were internally labelled to the dT of single-stranded DNA strands by means of a $C_6$ amino linker (modified by Integrated DNA Technologies, Inc.). A quartz microscope slide (Finkenbeiner) and coverslip were coated with polyethylene glycol (m-PEG-5000; Laysan Bio Inc.)[17,30] and biotinylated PEG (biotin-PEG-5000; Laysan Bio Inc.). Measurements were performed in a flow chamber that was assembled as follows. After the assembly of the coverslip and quartz slide[30], a syringe was attached to an outlet hole on the quartz slide through tubing. All the solution exchanges were performed by putting the solutions (0.1 ml) in a pipette tip and affixing it in the inlet hole, followed by pulling the syringe. The solutions were added in the following order. Neutravidin (0.2 mg $ml^{-1}$; Pierce) was applied to the surface and washed away with T50 buffer (10 mM Tris-HCl, pH 8, 50 mM NaCl). Biotinylated DNA (about 50–100 pM) in T50 buffer was added and washed away with imaging buffer (10 mM Tris-HCl, pH 8, 50 mM NaCl, 0.1 mg $ml^{-1}$ glucose oxidase, 0.02 mg $ml^{-1}$ catalase, 0.8% dextrose, plus Trolox)[31]. For replisome measurements, T7 gp4 (50 nM hexamer) and T7 DNA polymerase (gp5/thioredoxin) (50 nM) were loaded on the DNA with 2 mM dTTP, 5 mM DTT and 5 mM EDTA in imaging buffer, and incubated for 10 min. After a few seconds of imaging, unwinding and polymerase synthesis were initiated by the addition of the rest of the dNTPs (1 mM each), 1 mM ATP, 1 mM CTP, 5 mM DTT and 4 mM free $Mg^{2+}$ in imaging buffer. All measurements were made at room temperature.

FRET values were calculated as the ratio between the acceptor intensity and the total (acceptor plus donor) intensity after correcting for cross-talk between the donor and acceptor channels and subtracting the background. For the unwinding experiment shown in Fig. 1b, the initiation of FRET decrease and its saturation were scored by visual inspection of the donor and acceptor intensities and the calculated FRET efficiency, and the time difference between the two points was designated the $\Delta t$ value of each reaction. Once we had identifed a sustained FRET decrease below 0.5, the first time point at which FRET value dropped below the average FRET value before unwinding began, typically about 0.8, was designated the initiation point. Similarly, once we had identified a saturation in FRET decrease, the first time point at which FRET reached the average FRET value in the saturation plateau was designated the saturation point. We demonstrate that the error in determining the initial time point of FRET decrease by this method is negligible (Supplementary Fig. 5). The FRET increase time in the priming-loop experiment was plotted as the cumulative fraction against time to indicate the fraction of molecules that had completed the FRET increase up to a given time point. The donor intensity time was the time between the initial donor signal increase and the final disappearance of fluorescence signal, also determined by visual inspection. The donor intensity time was also plotted in the format cumulative fraction versus time. All data were analysed with scripts written in MATLAB and plotted in Origin.

31. Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nature Methods* **3**, 891–893 (2006).

ERRATUM

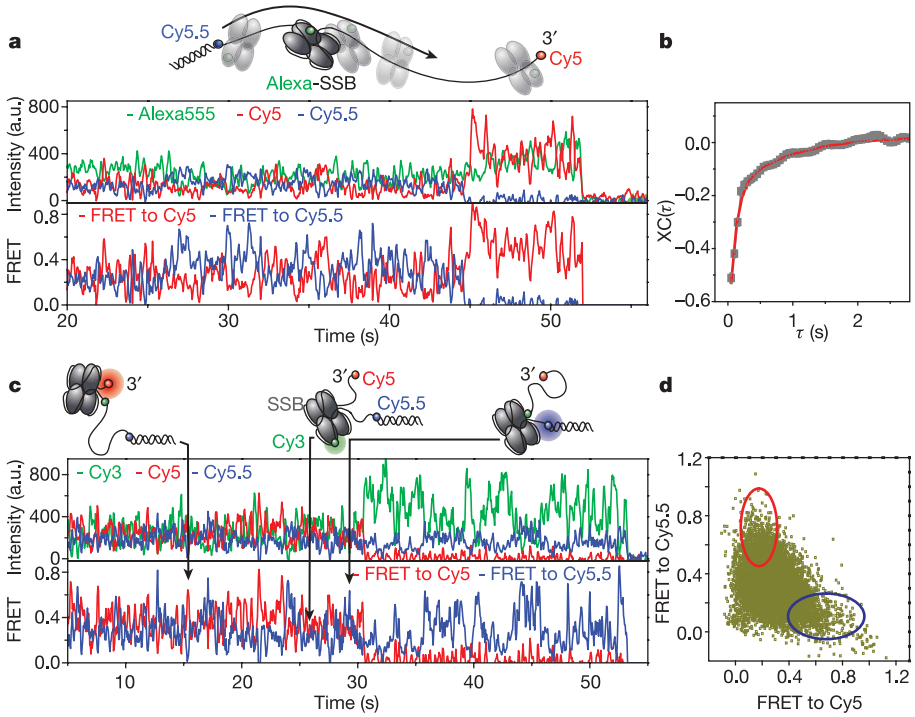
doi:10.1038/nature08600

# SSB protein diffusion on single-stranded DNA stimulates RecA filament formation

Rahul Roy, Alexander G. Kozlov, Timothy M. Lohman & Taekjip Ha

*Nature* **461**, 1092–1097 (2009)

In this Article, Figure 3 was printed incorrectly with missing labels. The correct figure is shown below.



NEWS

# Rekindling the gender-bias debate

It is almost five years since Harvard University's then-president Lawrence Summers made his controversial remarks about the innate differences between men's and women's skills in science and maths. Now, a book is reigniting the debate with equally contentious views about why female scientists have problems advancing in academia. *The Science on Women in Science* (AEI Press; 2009) is a collection of nine essays assembled and edited by former philosophy professor Christina Hoff Sommers. The book's aim, says Sommers, is to convince policy-makers and others that gender bias may not be the primary cause of the under-representation of women in science.

Six of the nine essays, including Sommers's own contribution, contend that innate gender differences, not bias, prompt men to be more interested than women in science, maths and engineering. Published by the American Enterprise Institute for Public Policy Research (AEI), a right-leaning think tank based in Washington DC, the book is Sommers's rejoinder to *Beyond Bias and Barriers*, a 2007 report by the US National Academy of Sciences, the National Academy of Engineering and the Institute of Medicine. That study found that under-representation of women in high-level academic posts is caused largely by bias and by an institutional framework that hinders women's advancement.

## Gender gap

In the collection's first essay, developmental psychopathologist Simon Baron-Cohen of the University of Cambridge, UK, suggests that men are more interested in analysing the variables in a system to determine its rules — or 'systemizing' — whereas women are more interested in empathizing. Baron-Cohen refers to psychological and behavioural studies that show that boys prefer certain toys and perform better than girls on some tests, and that men dominate occupations such as metalworking. These findings provide strong evidence, he writes, that men are more drawn to systems-based operations than women, and that women are more likely to prefer and be better at empathy-based activities than men. To highlight this, he notes that women's conversations involve more discussion about feelings compared with those of men.

But Elizabeth Spelke, a psychologist at Harvard University in Cambridge, Massachusetts, contends that the gender gap is due to societal and historical reasons. She cites behavioural, physiological and psychological studies suggesting that male and female infants are equally interested in a variety of objects (in contrast to the traditional view that girls are more interested in people than in objects); that girls and boys have an equal ability to learn the words that represent numbers and to count; and that girls can learn to read a map as well as boys can. "There is no evidence for a male advantage in intrinsic aptitude or motivation for mathematics and science," she writes. Spelke concludes that core cognitive and motivational patterns in women and men are much the same and that they play no part in the smaller numbers of women than men in science, maths and engineering.

IMAGES.COM/CORBIS

Contrary to the view of Sommers and others, Rosalind Chait Barnett, a senior scientist at the Women's Studies Research Center at Brandeis University in Waltham, Massachusetts, dismisses the notion that women prefer to avoid science and engineering because of their gender. In their essay, she and her co-author, social psychologist Laura Sabattini, argue that academia in the United States has historically been inhospitable to women scientists. College requirements of the mid-nineteenth century stipulated that women faculty members be single, they point out, and the twentieth century's anti-nepotism policies often meant a female scientist could not be employed at the same university as her husband. "Preferences are learned," says Barnett, noting that today, many women deliberately avoid the only occupations that employed them years ago. "The whole thing is completely culturally determined," she says.

Other studies have reached similar conclusions: Reshma Jagsi at the University of Michigan and her colleagues found that women are less likely than men to receive major funding for early-career clinical research (R. Jagsi *et al*. *Ann. Intern. Med*. **151,** 804–811; 2009). The authors blame family responsibilities and the excessive demands of clinical studies that take time away from research projects, suggesting that women are less successful than men at negotiating to reduce their clinical workload.

Sommers objects to initiatives such as the US National Science Foundation's ADVANCE programme, which seeks to increase women's representation and advancement in science, technology, engineering and maths. She also opposes the US government's application of 'Title IX' requirements — which bar gender discrimination in any federally funded operation — to academic science departments because, she believes, it could exclude male scientists in favour of females.

## Advancing the debate

In her essay, 'Sex, Science and the Economy', she argues that quota-driven and gender-balanced academic research is likely to be of lower quality than research conducted in departments without quotas. To support this, Sommers refers to studies suggesting that professions are partly biologically determined, and includes anecdotal evidence from female scientists who have not experienced bias. "Before the government rushes in to correct the problem, they have to be sure it exists and they also have to understand it. I'm trying to create a healthy scepticism," says Sommers. "I'm not absolutely convinced there isn't any bias," she says, but there are "more plausible explanations" for academic gender disparity.

But institutional bias has already been shown to be the main obstacle to women's ascension in academic science and engineering, says Mary Hall Reno, professor and chair of the department of physics and astronomy at the University of Iowa in Iowa City, and outgoing chair of the American Physical Society's Committee on the Status of Women in Physics. She gives the example of how bias was identified and addressed in the medical profession: "Forty years ago, the thought was that women couldn't be medical doctors, and now they have thronged into the field," Reno says. "I thought we were past this. We don't need more debate."

Phoebe Leboy, president of the Association for Women in Science in Washington DC, says that the book "cannot be ignored", noting its potential to prompt the overturn of anti-discrimination legislation or programmes such as ADVANCE if lawmakers were to act on its conclusions. But she thinks that scenario is unlikely. "This book will preach mostly to the converted, and the converted are a relatively small group of people with unusually good access to the media," she says. ■

**Karen Kaplan**

NEWS

# Japanese mentors reap their awards

The winners of the sixth annual Nature Awards for Mentoring in Science, announced on 1 December, have had very different career paths. One has been doing the rounds at Japan's elite academic institutions for decades; the other is a slightly eccentric product of industrial laboratories. But they share scientific initiative as well as an open-mindedness to their students' ideas and a willingness to foster their students' independence.

**Winning mentors: Hiroaki Kitano (left) and Fumio Oosawa.**

The lifetime achievement award went to biophysicist Fumio Oosawa, who pioneered molecular studies on muscle contraction and helped invent the field of single-molecule biology. The midterm career award was won by computer scientist Hiroaki Kitano, who is renowned for his expertise in robotics and systems biology. The prizes, which were presented by Philip Campbell, editor-in-chief of *Nature*, at a reception in Tokyo, each carried a purse of ¥1.5 million (US$17,000).

Like all the winners of *Nature*'s mentor awards, Oosawa and Kitano share certain characteristics, says Campbell. They have broad vision, a keen interest in challenging students and a willingness to help them network to advance their careers.

Oosawa received his physics degree in 1944 from the University of Tokyo before moving on to Nagoya University and then Osaka University. Now 77, he initiated his research on muscle contraction in the 1960s by analysing the polymerization of actin. Kitano, 48, heads Sony Computer Science Laboratories in Tokyo. He began his career with information-technology company NEC in Tokyo.

## Great support

Testimony from current and former students, who nominated the pair for the awards, reveals some of the characteristics shared by superior mentors. Kitano and Oosawa both cultivate independence in their researchers. Kitano, for example, routinely names laboratory members as corresponding authors on published papers. Both were praised for using their networks of colleagues and contacts to introduce their lab members to other established researchers in the field. "A lot of professors just don't take the time to do that," says Samik Ghosh, a researcher at the Systems Biology Institute in Tokyo, headed by Kitano.

Speaking at the awards ceremony, Akiyoshi Wada, who headed the judging committee, said: "It's obvious that one generation of scientists should cultivate the next. But it's too obvious — so obvious that people take it for granted." Oosawa and Kitano did not, Wada added.

Kitano and Oosawa share a broad-minded approach to ideas that stimulate their students. "He takes an interest in every project," wrote one of Oosawa's nominators. In a written statement, Oosawa said that he makes sure to "respect the ideas of my students. I never said things like, 'I already thought of that'."

## Absurd ideas

One student who nominated Kitano wrote: "He is ready to invest wholeheartedly in absurd ideas, and in some cases it seemed the more absurd the better." Kitano's 'absurd' ideas spawned RoboCup — a seminal competition that aims to produce robots, currently competing against each other, that can take on a human team — and the idea of a 'virtual human' — a computer model of a person, used in drug development (see *Nature* **451,** 879; 2008).

Such radical ideas appeal to young Japanese researchers. Hiroki Ueda of the Center for Developmental Biology, a former student in Kitano's laboratory and one of Japan's most prodigious talents, plays down the importance of the science knowledge he gleaned from working with Kitano. "I learned nothing in his laboratory," he says playfully. But Ueda, who considers himself a radical like Kitano, says he picked up "the atmosphere", which helped convince him that a career can be founded on unconventional ideas. "That's so cool," Kitano said when told of Ueda's comment.

In the laboratory, students address Oosawa with the more-familiar honorific suffix '*san*' rather than the formal '*sensei*'. "He had little time for authoritarianism," says one student. Instead of an office, he had a table at the end of the laboratory, but he rarely used it, preferring to walk about and talk to students. Instead of the Oosawa 'lab', students spoke of a sort of nurturing open classroom where students could come and go with ease. Oosawa says he was careful not to push for a debate about findings while a student was still in the middle of collecting data. "I wait," he says. "I don't judge by the speed of the work. Having interesting ideas is the most important thing." Breaking down barriers made Oosawa an early force in equal rights: in the 1960s, at a time when few Japanese women could find a career in science, a third of his laboratory was female.

## Tough environment

The accomplishments of Oosawa and Kitano are all the more impressive considering their research environment. In Japan, a strong mentor–protégé culture often promotes deference rather than independence.

There are also systemic issues that can hinder the proliferation of great mentors, says Hitoshi Murayama, a theoretical physicist with positions at the University of Tokyo, the Lawrence Berkeley National Laboratory in California and the University of California, Berkeley. Japanese universities generally do not provide stipends to graduate students, so many have to work part-time to get through school. In Japan, professors can take anyone who comes, without necessarily worrying about the quality or future of their students. By contrast, US professors generally invest in young researchers by providing financial support from their grant monies, and so have more at stake. "Professors don't take on students unless they are serious about it," says Murayama, noting an upside to the US system. "This definitely creates a difference in the mentor–student relationship."

Oosawa and Kitano, though, have managed to thrive, and to help their students do the same — and both their own and their students' science careers have benefited as a result. As Campbell noted at the awards ceremony: "Not every outstanding scientist is a good mentor, but all outstanding mentors are outstanding scientists." ■

**David Cyranoski**

**See Editorial, page 826.**

# Rejuvenation

The chase is on.

Julian Tang

"So, can anyone tell me about the Hoffman vortex?"

Mr Murphy looked expectantly at his small A-level physics class as some clusters of excited murmuring broke out.

"Conferring is allowed," he chuckled. "Catherine, Rebecca, come on! Tell me what you think before these rowdy boys beat you to it!"

Looking at her and grinning, Rebecca nudged Catherine, urging her to speak first.

"Well," Catherine began hesitantly, as the boys looked up, curiously. "Discovered by the physicist, Hans Hoffman, it was found that when two rings of a specific diameter and thickness of the radioactive element zylerium are counter-rotated at a specific rate about the same axis, at a specific distance apart, a vortex of zyleron particles is produced between the two rings."

Rebecca clapped her on the back, congratulating her on her answer.

"Perfect!" exclaimed Mr Murphy in delight. "Now boys, what does this vortex do? Come on now, don't let the girls show you up!"

"Well, it was only one girl, sir," retorted Paul, leader of one of the boys' 'think tank' clusters.

"True, I stand corrected," replied Mr Murphy, in mock humility.

Jonathan, the leader of one of the other boy think tanks, raised his right hand.

"Ah! Jonathan! What can you offer me on this fine day?" Mr Murphy looked at him, hopefully.

Jonathan scratched his right ear, a habit of his before answering a question in class. "The Hoffman vortex created by two rings of zylerium, arranged in the way that Catherine described, has been postulated to reverse the effects of ageing, mainly by reversing the effects of solar radiation and other ageing processes at a cellular level."

"Yes, well done! We're firing on all thrusters today, class!" Mr Murphy grinned widely and clapped his hands, as Jonathan's friends punched him, playfully.

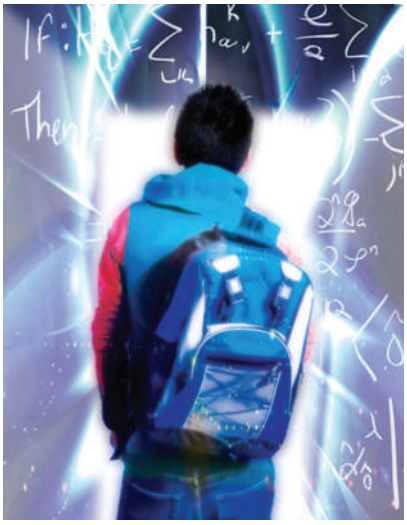
"OK, now Paul, have you anything else to add?" Mr Murphy clasped his hands on his stomach and twiddled his thumbs in expectation.

After some brief, last-minute conferring with his friends, Paul cleared his throat loudly (something that he always did before speaking in front of the class).

"Yes, all of that is fine, but these effects have only been shown in microorganisms, such as bacteria. The problem is that the amount of zylerium we can produce and purify at this time is in the order of milligrams. To make zylerium rings large enough to produce a man-sized Hoffman vortex will need a much more efficient way to produce zylerium," he finished proudly.

"Excellent!" said Mr Murphy, bowing slightly in Paul's direction.

Then another student, Ben, put his hand up. He was relatively new, having joined the class at the beginning of the current school-term. Quietly assertive and intelligent, he seemed to be mature beyond



his teenage years. He had quickly gained the respect (and in some cases, awe) of his classmates and teachers.

"Yes, Ben," nodded Mr Murphy, with interest.

"There is another hypothesis, sir," he started, carefully. "The addition of a third zylerium ring creates a modified Hoffman vortex, which allows space-time travel, on top of the rejuvenation effect." Ben smiled, knowingly, at him.

All the students looked at him in amazement.

"Where did you hear that, Ben?" blurted out Paul. "Has it been published somewhere, yet?"

Ben shook his head slightly. "Let's just say that this was a 'personal communication' of some sort," replied Ben, evasively. He was still looking at Mr Murphy, intently.

Mr Murphy returned his look, impassively for a moment then glanced around, quickly. "OK class, you can all leave a bit earlier today. Ben, can you stay behind for a minute?"

Ben nodded, still smiling at him and remained seated.

Bemused, the rest of the students filed out in silence, leaving the two of them alone.

Mr Murphy shut the door then sat on the front bench, facing Ben.

"So, well done Ben, or shall I call you by your real name, Jan Bendarian, the bounty hunter? The rejuvenation process suits you," Mr Murphy sneered.

Ben/Jan did not move from his seat. "We don't care about you, Murphy, we just want the device. What are you planning to do? Reduce this world's population to a bunch of babies and let them starve to death at your whim?"

"What? Me? Be so inhumane? Never!" Murphy seemed sincere in this. "This world has treated me well. It accepted me as one of them, even letting me teach their young minds. I had planned to settle here into my old age, peacefully. Now, you've spoiled all that," he finished, grimly.

While he was saying this, Murphy had casually moved back behind the bench, in front of the blackboard. Now, suddenly, he hit a hidden switch under the bench-top then, waving and laughing at Ben/Jan, he walked backwards *through* the blackboard and wall behind him, disappearing. The blackboard and wall reformed after him.

Ben/Jan, still seated, sighed and took out a small device from his pocket. Yes, the microscopic tracking beacon he had slipped into Murphy's coffee last week was still functioning. It would have lodged somewhere in his gut, where it would remain, unless Murphy had it surgically removed.

Examining the bench and blackboard, he found the switch easily and entered Murphy's inter-dimensional storage area behind. Murphy had long gone to some other space-time destination, probably already with a different age and disguise, with the help of the Hoffman vortex generator.

What the hell, he thought. He had already pursued him through five locations, what was one more? He always enjoyed a good chase. Chuckling to himself, he turned on his belt-mounted Hoffman field (a more advanced, portable form of the Hoffman generator) whereupon his image shimmered briefly, then he too disappeared. ■

JACEY

**Julian Tang is a clinical/academic virologist. This story is dedicated to his real 'A'-level physics class led by the real Mr Murphy, who kept us all stimulated with his humour and sarcastic wit.**

**Join the discussion of Futures in *Nature* at go.nature.com/QMAm2a**